Unsupervised Exceptional Attributed Sub-graph Mining in Urban Data

Anes Bendimerad, Marc Plantevit, Céline Robardet

Unsupervised Exceptional Attributed Sub-graph Mining in Urban Data

Ahmed Anes Bendimerad*, Marc Plantevit†, Céline Robardet*
*INSA de Lyon, CNRS, LIRIS UMR5205, F-69621 France
†Université Lyon 1, CNRS, LIRIS UMR5205, F-69622 France
{firstname.name}@liris.cnrs.fr

***Abstract*—Geo-located social media provide a wealth of information that describes urban areas based on user descriptions and comments. Such data makes possible to identify meaningful city neighborhoods on the basis of the footprints left by a large and diverse population that uses this type of media. In this paper, we present some methods to exhibit the predominant activities and their associated urban areas to automatically describe a whole city. Based on a suitable attributed graph model, our approach identifies neighborhoods with homogeneous and exceptional characteristics. We introduce the novel problem of exceptional sub-graph mining in attributed graphs and propose a complete algorithm that takes benefits from new upper bounds and pruning properties. We also propose an approach to sample the space of exceptional sub-graphs within a given time-budget. Experiments performed on 10 real datasets are reported and demonstrate the relevancy and the limits of both approaches.**

I. Introduction

In today's increasingly global and interconnected world, people have opportunities to live abroad of their country, generally in urban areas. They face the challenge of making decisions about where to live, how to find appropriate areas to go out or a place to visit. Thanks to the current numerical development, numerous sources of collected data can help to make better decisions. Nevertheless, such geo-enabled social data must be processed with efficient methods to take into account the heterogeneity and the complexity of urban areas by the discovery of useful and understandable insights. Such questions have recently raised the interests of researchers such as discovering similar neighborhoods across several cities [5] or matching social attributes with geographic spaces [19].

Using social and urban data of a city (such as the ones provided by social networks as FOURSQUARE or GOOGLE-PLACE), we aim to identify neighborhoods with homogeneous and exceptional characteristics: Areas are described by their associated characteristics that distinguish them from the rest of the city. To this end, we propose a suitable attributed graph model (as illustrated in Fig. 1) that results from the combination of social and urban data, and we achieve the task by applying a constraint-based graph pattern mining approach. The devised algorithm identifies connected sub-graphs associated to some characteristics that discriminate the subgraphs from the rest of the graph.

Attributed graph analysis has received much attention in the past decade. For example, [14] designed a method to find dense homogeneous sub-graphs, where vertices are described by categorical attributes and [7] proposes subspace clustering approach using numerical vertex attributes. However, all these works focus their attention on the similarity inside the sub-graphs, while underestimate exceptionality of the sub-graph characteristics with respect to the whole graph.

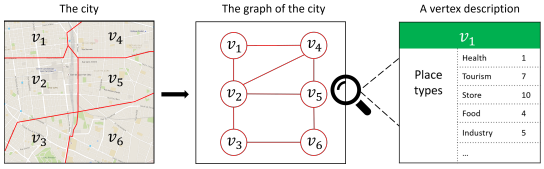


Fig. 1: Example of a graph modeling a city.

We design two algorithms to solve this problem. The first one is an exact algorithm that uses original and efficient upper bounds and some other techniques to reduce the search space. We also propose a method that reduces the number of output patterns by providing a concise summary of the complete result set while limiting the overlap between patterns. This summary represents the whole set of solutions, while being faster to compute. The second algorithm mines exceptional sub-graphs by sampling the space of patterns in a similar way as [2], [8], [17].

Our main contributions are manifold: First, we propose a new kind of graph analysis that exploits both of the contrasts of vertices attributes and the graph structure with a connectivity constraint. Second, to solve the defined problem, we present an efficient algorithm based on new upper bounds and pruning properties. We also propose a method to summarize the output set. Third, we design a probabilistic approach that samples the output space of patterns within a time budget specified by end-users. Forth, we provide a thorough empirical study that includes (1) a demonstration of the efficiency of the used pruning techniques, (2) the impact of the parameters and the input graph dimensions on the performance of the algorithms, and (3) the relevancy of the discovered results.

The rest of the paper is organized as follows. Section II formally introduces the problem. The proposed solutions are presented in Section III. Section IV reports on the empirical study on both synthetic and real-world datasets. Section V discusses related work and section VI concludes.

II. Problem formulation

Data describing geographic venues are numerous, ranging from census data to collaborative data produced through social-media platforms. To describe a city, nearby venues are grouped into small areas (geographers generally use tiles of 200 meters) over which venue characteristics are aggregated into count data. These areas are hereafter considered as the vertices V of a graph $G = (V, E, C, D)$ whose edges E connect adjacent areas (that share a part of their borders), $C = \{c_i,\ i \in [\![1, p]\!]\}$ is a set of p categories and the vertices of V are described by $D = \{c_i(v) \in \mathbb{N},\ \text{with } c_i \in C \text{ and } v \in V\}$, the counts of venues of each category in the area associated to each vertex.

The values of D can be aggregated over a set of vertices $K \subseteq V$ and a set of categories $L \subseteq C$: $sum(L, K) = \sum_{v \in K} \sum_{c_i \in L} c_i(v)$. To simplify the notation, we use $sum(K)$ to denote $sum(C, K)$.

As an example, Fig. 1 presents a graph derived from the division of a city into 6 areas (from v_1 to v_6). The area represented by v_1 is adjacent to the ones represented by v_2 and v_4, and consequently an edge connects v_1 to v_2 and another one v_1 to v_4. The number of venues of each category in a given area composed a vector associated to the corresponding vertex. The distribution of venue categories $C = (Health, Tourism, Store, Food)$ is detailed in Fig. 2. $sum(\textit{health}, \{v_1\}) = 1$ as there is one venue with the category *health* in the area associated to v_1. We can also observe that $sum(\{Health, Tourism, Store, Food\}, \{v_1\}) = 22$, and for the set $K = \{v_2, v_5\}$, $sum(K) = 49$.

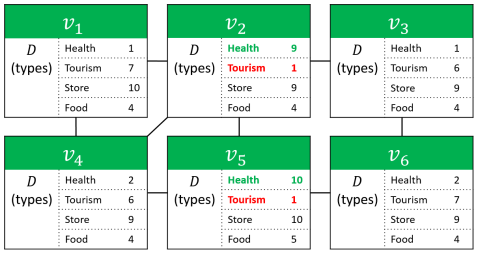


Fig. 2: Example of the distribution of venues in areas.

Our objective is to identify neighborhoods whose characteristics distinguish them from the rest of the city. Therefore, we propose to discover connected sub-graphs associated to exceptional categories. A category is exceptional for a sub-graph if it is more frequent in its vertices than in the remaining of the graph. The scarcity of a category can also be a relevant descriptive element of a neighborhood. In Fig. 2, vertices v_2 and v_5 have a surplus on the category *Health* compared to the rest of the graph, while having a loss on category *Tourism*. Thus, we associate to a sub-graph a characteristic defined as a pair $S = (S^+, S^-)$, with S^+ and S^- two disjoint subsets of C. In order to assess the relevancy of the characteristics S with respect to the sub-graph induced by $K \subseteq V$, noted $G(K)$, we define the measure $WRAcc(S, K)$, an adaptation of the weighted relative accuracy measure widely used in Subgroup Discovery [9].

A set of categories L is discriminant to $G(K)$ if it is more or, on the contrary, less frequent in $G(K)$ than in G. This is evaluated by the $gain$ function:

$$gain(L, K) = \frac{sum(L, K)}{sum(K)} - \frac{sum(L, V)}{sum(V)}$$

The validity of a characteristic $S = (S^+, S^-)$ with respect to $G(K)$ is given by

$$valid(S, K) \equiv \bigwedge_{v \in K} \Big(\big(\bigwedge_{c_i \in S^+} \delta_{gain(c_i, v) > 0} \big) \bigwedge \big(\bigwedge_{c_i \in S^-} \delta_{gain(c_i, v) < 0} \big) \Big)$$

$valid(S, K)$ means that each vertex $v \in K$ has a positive gain for each category $c_i \in S^+$, and a negative gain for each category $c_i \in S^-$. The quality of characteristic S can be globally measured by the numerical function A:

$$A(S, K) = gain(S^+, K) - gain(S^-, K)$$

However, a major drawback of the gain is that it is easy to obtain high value with highly specific characteristics [9], more precisely characteristics associated to a small set of vertices. Weighted relative accuracy makes a trade-off between generality and gain by considering the relative size of the sub-graph:

$$WRAcc(S, K) = \begin{cases} A(S, K) \times \frac{|K|}{|V|} \text{ if } valid(S, K) \\ 0 \text{ otherwise} \end{cases}$$

The main differences with the $WRAcc$ used in Subgroup Discovery [9] are (1) our adapted $WRAcc$ considers both the positive and the negative contrasts in an unsupervised setting, (2) it takes into account the homogeneity of elements of K, using the predicate $valid(S, K)$. We can now define the pattern domain we consider:

Definition 1 (Exceptional sub-graph)**.** *Given a graph $G = (V, E, C, D)$ and two thresholds σ and δ, an exceptional sub-graph (S, K) is such that (1) $|K| \geq \sigma$, (2) $G(K)$ is connected, and (3) $WRAcc(S, K) \geq \delta$. $\mathcal{E}(G)$ denotes the whole set of exceptional sub-graphs in G.*

The computation of the complete set of exceptional sub-graphs requires to search into two combinatorial search spaces, with constraints that cannot be used according to the usual techniques of search space pruning. Thus, a naive approach cannot achieve this task for large graphs or large number of categories. In the following we propose several enumeration strategies that takes benefit from computation of tight upper-bounds on $WRAcc(S, K)$.

III. Computing exceptional sub-graphs

This section introduces three distinct approaches to extract exceptional sub-graphs. First, we present an exact algorithm that aims at discovering the complete set of exceptional sub-graphs. Second, we propose to only compute a summary of this set by controlling the overlap between patterns. Third, we devise a heuristic algorithm that samples the space of

exceptional sub-graphs within a user-defined time-budget. This approach makes possible to obtain instant results and to successfully scales up to a large number of attributes.

A. The complete approach

To compute all exceptional sub-graphs, two search spaces have to be explored: The space of characteristics $S = (S^+, S^-)$ is first traveled, and then, for each promising characteristic, the sub-graphs $G(K)$, $K \subseteq V$, are considered.

1) Characteristic enumeration: We explore the search space of characteristics recursively. We start from the empty characteristic $(S^+, S^-) = (\emptyset, \emptyset)$ and consider the candidate categories that can be used to expand S: $(X^+, X^-) = (C, C)$. X^+ contains the categories that can be added to S^+, and X^- the ones that can be added to S^-.

However, if the size of the category set C is large, it is unrealistic to explore all the combinations. To early discard unpromising candidates, we exploit two properties: The anti-monotony of the predicate *valid* and an upper-bound on the $WRAcc$ measure.

The predicate *valid* is anti-monotone with respect to the inclusion of characteristics: Considering two characteristics S_1, S_2 such that $S_1^+ \subseteq S_2^+$ and $S_1^- \subseteq S_2^-$, denoted $S_1 \subseteq S_2$, we have $\forall K \subseteq V$, $valid(S_2, K) \Rightarrow valid(S_1, K)$. Let $V(S) = \{v \in V \mid valid(S, v)\}$, if $V(S)$ has a size smaller than σ, then it is unnecessary to explore the children of S. To quickly obtain a small-sized set $V(S)$, we expand S^+ with the candidate $c \in X^+$ that has the smallest number of vertices $v \in V$ with $gain(c, v) > 0$ (*a.k.a.* fail-first principle), and to expand S^- with the candidate $c \in X^-$ that has similar property for $gain(c, v) < 0$ (lines 1 and 2 of Algorithm 1).

Let us now define an upper bound on the $WRAcc$ measure. Let $\Omega(S, X)$ be the exploration tree rooted in (S, X) that contains all the characteristics T obtained from the characteristic S extended with some categories of X. An upper bound UB_1 on $WRAcc$ with respect to a connected component $CC \subseteq V(S)$ is defined as: $UB_1(S, X, CC) = \max_{S \in \Omega(S,X), v \in CC} A(S, v) \times \frac{|CC|}{|V|}$.

Property 1. *$WRAcc(T, K) \leq UB_1(S, X, CC)$, $\forall T \in \Omega(S, X)$ and $K \subseteq CC$*

Proof. From the definition of the function A, we have $A(S, K) = \sum_{v \in K} \frac{sum(v)}{sum(K)} \times A(S, v)$. Thus, $A(S, K)$ is simply a weighted average of $A(S, v)$. Furthermore, as $A(S, v) \leq A(S^\star, v)$ with $S^\star = \text{argmax}_{S \in \Omega(S,X)} A(S, v)$, from the properties of the average, we can infer that $A(S, K) \leq A(S^\star, v)$. Thus, we can conclude that $A(S, K)\frac{|K|}{|V|} \leq A(S^\star, v)\frac{|CC|}{|V|}$. □

The computation of UB_1 can be done in $\Theta(|CC| \times |X|)$. Algorithm 1 presents the pseudo-code of ENERGETICS[1] that computes the exceptional sub-graphs. Line 3 stops the algorithm when there is no more candidate. The loop in line 4 discards unpromising connected components using the aforementioned upper bound. The four recursive calls consider the four possible extensions of S by a category c. The function SUB-GRAPH is detailed below.

Algorithm 1: ENERGETICS(S, X, Y, R, δ, σ)

Input: $S = (S^+, S^-)$ the current explored characteristic, $X = (X^+, X^-)$ the candidate sets, $Y = \{v \in V \mid valid(S, v)\}$

Output: R the result set under construction

1 Sort $c \in X^+$ in ascending order of $|\{v \in V \mid gain(c, v) > 0\}|$
2 Sort $c \in X^-$ in ascending order of $|\{v \in V \mid gain(c, v) < 0\}|$
3 **if** $X \neq (\emptyset, \emptyset)$ **then**
4 **for** $CC \subseteq G(Y)$ **do**
5 **if** $UB_1(S, X, CC) < \delta$ *or* $|CC| < \sigma$ **then**
6 $Y \leftarrow Y \setminus CC$
7 **if** $X^+ \neq \emptyset$ **then**
8 $c \leftarrow \text{pop}(X^+)$
9 $Y' \leftarrow \{v \in Y \mid valid((S^+ \cup \{c\}, S^-), v)\}$
10 **if** $Y' \neq \emptyset$ **then**
11 ENERGETICS($(S^+ \cup \{c\}, S^-)$, $(X^+ \setminus \{c\}, X^- \setminus \{c\})$, Y', R, δ, σ)
12 SUB-GRAPH($(S^+ \cup \{c\}, S^-)$, Y', δ, σ, R)
13 ENERGETICS(S,$(X^+ \setminus \{c\}, X^-)$,$Y$,$R$, δ, σ)
14 **else**
15 $c \leftarrow \text{pop}(X^-)$
16 $Y' \leftarrow \{v \in Y \mid valid((S^+, S^- \cup \{c\}), v)\}$
17 **if** $Y' \neq \emptyset$ **then**
18 ENERGETICS($(S^+, S^- \cup \{c\})$, $(X^+, X^- \setminus \{c\})$, Y', R, δ, σ)
19 SUB-GRAPH($(S^+, S^- \cup \{c\})$, Y', δ, σ, R)
20 ENERGETICS(S,$(X^+, X^- \setminus \{c\})$,$Y$, R, δ, σ)

2) Sub-graph exploration: Once the characteristic S is fixed, its associated set of vertices $V(S)$ has to be processed to identify subsets K' that (1) have a size greater or equal to σ, (2) satisfy $WRAcc(S, K') \geq \delta$ and (3) K' are connected. To compute them efficiently, we propose an approach that combines reverse search [20] and upper-bounds. It is based on the loose anti-monotonicity of the constraint $A(S, K) \geq \tau$.

Property 2 (Loose anti-monotone)**.** *$A(S, K) \geq \tau$ is a loose anti-monotone constraint. This implies that if $A(S, K) \geq \tau$, then it exists $v \in K$ such that $A(S, K \setminus \{v\}) \geq \tau$.*

Proof. Let $v^\circ = \text{argmin}_{u \in K} A(S, u)$ and $K = K' \cup \{v^\circ\}$. By construction, $\forall u \in K$, $A(S, u) \geq A(S, v^\circ)$. Thus, the weighted average $\frac{\sum_{u \in K'} sum(u) A(S,u)}{\sum_{u \in K'} sum(u)} \geq A(S, v^\circ)$ that is to say $A(S, K') \geq A(S, v^\circ)$. As $A(S, K) = \sum_{v \in K} \frac{sum(v)}{sum(K)} A(S, v)$, we can derive that $A(S, K) = \frac{sum(K')}{sum(K)} A(S, K') + \frac{sum(v^\circ)}{sum(K)} A(S, v^\circ) \leq A(S, K') \frac{sum(K') + sum(v^\circ)}{sum(K)} = A(S, K') = A(S, K \setminus \{v^\circ\})$. □

To exploit this property, we enumerate the vertex sets K in a way such that $A(S, K)$ decreases during the enumeration. Starting from a sub-graph $K = \emptyset$ and a set of candidates vertices Y, K is expanded by adding vertices from Y according to the following order.

Definition 2 (Vertex order $\succeq_K$)**.** *Let $K \subseteq V$ and $v_i, v_j \in V \setminus K$. We say that $v_i \succeq_K v_j$ iff $A(S, K \cup \{v_i\}) \geq A(S, K \cup \{v_j\})$*

[1] ENERGETICS: ENumERatinG ExcepTIonal Connected Sub-graphs

From the loose anti-monotonicity of $A(S,K)$, we can derive an upper bound of the $WRAcc$ constraint.

Definition 3 (UB_2)**.** *Let K be the current set of vertices and Y the set of candidates. The function UB_2 is defined as:*

$$UB_2(S,K,Y) = A(S,K) \times \frac{|K \cup Y|}{|V|}$$

Property 3. *For all $G(K')$ with $K' = K \cup K''$ and $K'' \subseteq Y$, we have*

$$WRAcc(S,K') \leq UB_2(S,K,Y)$$

Proof. Let $K' = K \cup K''$. By generalizing the result of proof 2, we have $A(S,K') \leq A(S,K)$. Moreover, as $K' \subseteq K \cup Y$, we have $\frac{|K'|}{|V|} \leq \frac{|K \cup Y|}{|V|}$. By multiplying these results we can conclude the proof. □

Therefore, if $UB_2(S,K,Y) < \delta$, we are sure that there is no set $K' \subseteq K \cup Y$ such that $WRAcc(S,K') \geq \delta$ and we can stop the enumeration.

We can derive a tighter upper-bound constructed recursively on the size of the set of vertices $Z \subseteq Y$ added to K in the enumeration process. Considering the vertex order $\succeq_K$ used during the enumeration, we have $A(S,K) \geq A(S,v),\ \forall v \in Y$. Thus, the second term in the following weighthed average $A(S,K \cup \{v\}) = \frac{sum(K)}{sum(K)+sum(v)} A(S,K) + \frac{sum(v)}{sum(K)+sum(v)} A(S,v)$ tends to decrease $A(S,K \cup \{v\})$. To tightly upper bound $A(S,K \cup \{v\})$, we have to give a high value to $\frac{sum(K)}{sum(K)+sum(v)}$, while minimizing $\frac{sum(v)}{sum(K)+sum(v)}$. This is achieved by $d_1 = \min_{v \in Y} sum(v)$:

$$A(S,K \cup \{v\}) \leq \frac{sum(K)}{sum(K)+d_1} A(S,K) + \frac{d_1}{sum(K)+d_1} A(S,v)$$

Then, by choosing $a_1 = \max_{v \in Y} A(S,v)$, the above expression is upper bounded. We just have defined $\overline{A}(S,K,Y,1)$, the upper bound for K extended by a single vertex from Y.

Let us now suppose there is an upper bound $\overline{A}(S,K,Y,i)$ of $A(S,K \cup Z)$ with $|Z| = i$. To define $\overline{A}(S,K,Y,i+1)$ for $|Z| = i+1$, we consider $A(S,K \cup Z \cup \{v\}) = \frac{sum(K)+sum(Z)}{sum(K)+sum(Z)+sum(v)} A(S,K \cup Z) + \frac{sum(v)}{sum(K)+sum(Z)+sum(v)} A(S,v)$. $A(S,v)$ tends to decreases the weighted sum, thus (1) $\frac{sum(K)+sum(Z)}{sum(K)+sum(Z)+sum(v)}$ has to be maximized and (2) $\frac{sum(v)}{sum(K)+sum(Z)+sum(v)}$ has to be minimized. Let $|Y| = t$ and $\{d_1,\ldots,d_t\} = \{sum(v) \mid v \in Y\}$ such that $\forall i,j \in [\![1,t]\!],\ d_i \leq d_j$. (1) is obtained by $\frac{sum(K)+\sum_{j=t-i+1}^{t} d_j}{sum(K)+\sum_{j=t-i+1}^{t} d_j+d_1}$, and (2) by $\frac{d_1}{sum(K)+\sum_{j=t-i+1}^{t} d_j+d_1}$. The upper bound is achieved by replacing $A(S,v)$ by $a_{i+1} = \max_{v \in Y \setminus Z} A(S,v)$ and $A(S,K \cup Z)$ by $\overline{A}(S,K,i)$. This is synthesized in the following definition.

Definition 4 (UB_3)**.** *Let (K,Y) be the current enumerated sub-graph induced by K, and Y the candidate vertices such that $|Y| = t$. Let $\{d_1,\ldots,d_t\} = \{sum(v) \mid v \in Y\}$ such that $\forall i,j \in [\![1,t]\!]$ with $i \leq j$, $d_i \leq d_j$, and $\{a_1,\ldots,a_t\} = \{A(S,v) \mid v \in Y\}$ such that $\forall i,j \in [\![1,t]\!]$ with $i \leq j$, $a_i \geq a_j$. The upper bound UB_3 is*

$$\begin{aligned}
UB_3(S,K,Y) &= \max_{i \in [\![0,t]\!]} \left(\bar{A}(S,K,Y,i) \times \tfrac{|K|+i}{|V|} \right) \text{ with} \\
\bar{A}(S,K,Y,0) &= A(S,K) \\
\bar{A}(S,K,Y,i) &= \tfrac{sum(K)+\sum_{j=t-i+2}^{t} d_j}{sum(K)+\sum_{j=t-i+2}^{t} d_j+d_1} \bar{A}(S,K,Y,i-1) \\
&\quad + \tfrac{d_1}{sum(K)+\sum_{j=t-i+2}^{t} d_j+d_1} \times a_i,\ i \geq 1
\end{aligned}$$

A last pruning technique is obtained by observing that if a branch of the sub-graph enumeration tree does not lead to any valid pattern, then its right sibling cannot neither contain any valid patterns.

Property 4 (Sibling-based upper bound)**.** *In the enumeration tree, let $\Gamma(K,Y)$ be the set of $K' \subseteq K \cup Y$ such that $WRAcc(S,K') \geq \delta$ and K' is generated from (K,Y). Let $v_1, v_2 \in Y$ be such that $v_1 \succeq_K v_2$ and $A(S,v_1) \geq A(S,v_2)$. If $\Gamma(K \cup \{v_1\}, Y \setminus \{v_1\})$ is empty then $\Gamma(K \cup \{v_2\}, Y \setminus \{v_1,v_2\})$ is also empty.*

Proof sketch. We prove the contraposed: $\Gamma(K_2,Y_2) \neq \emptyset \Rightarrow \Gamma(K_1,Y_1) \neq \emptyset$, with $K_1 = K \cup \{v_1\}$ and $K_2 = K \cup \{v_2\}$. By considering a valid pattern $F \in \Gamma(K_2,Y_2)$, we derive another pattern by removing from F the vertex $v^\circ = \text{argmin}_{u \in F \cap Y} A(S,u)$ and adding the vertex v_1. This pattern (1) has the same size than F, (2) belongs to $\Gamma(K_1,Y_1)$ and (3) is such that $A(S,F \setminus \{v^\circ\} \cup \{v_1\}) \geq A(S,F)$. This concludes the sketch of the proof. □

Algorithm 2 presents the implementation of SUB-GRAPH that calls the function SUB-CC on each maximal connected component CC of $G(V(S))$. SUB-CC returns TRUE if it finds in the explored sub-tree a sub-graph $G(K)$ such that $WRAcc(S,K) \geq \delta$. When FALSE, the set of candidate vertices is reduced based on property 4. The algorithm SUB-CC begins by checking the upper bounds UB_2 and UB_3 (lines 4 and 6). Then, in the while loop (line 9) it enumerates a new candidate that is recursively expanded through the recursive call of SUB-CC. If the last call in line 12 returns FALSE, which means that $\Gamma(K \cup \{v^\star\}, Y) = \emptyset$, all the vertices v such that $A(S,v) \leq A(S,v^\star)$ are pruned (line 15).

B. Reducing the redundancy among patterns

For the same characteristic $S = (S^+,S^-)$, $\mathcal{E}(G)$ can contain several patterns which may be very similar. This has two main disadvantages: (1) the size of the result set may be uselessly very large and redundant, and (2) the method performance may degrade due to the size of the output.

We tackle this problem in a similar way as [22] that proposes an approach to reduce the output set of maximal cliques. The idea consists in returning a concise subset of $\mathcal{E}(G)$, denoted $\mathcal{E}'(G)$, that represents the whole set of patterns. More precisely, for each pattern $(S,K_1) \in \mathcal{E}(G)$ there is a pattern $(S,K_2) \in \mathcal{E}'(G)$ such that K_1 is included in K_2 or K_2 covers the majority of K_1. This is formally defined below.

Definition 5 (Coverage measure)**.** *Let $K_1 \subseteq V$ and $K_2 \subseteq V$. The coverage measure cov is defined by $cov(K_1,K_2) = \frac{|K_1 \cap K_2|}{|K_1|}$. This function measures how much K_2 covers K_1.*

Algorithm 2: SUB-GRAPH(S,$validV$,δ,σ,R)

Input: S, $validV = \{v \in V \mid valid(S, v)\}$
Output: R the set of result patterns under construction
for ***each*** *maximal connected component* $CC \subseteq validV$ **do**
  SUB-CC(S,($\emptyset, CC$),R,δ,σ)

SUB-CC(S,(K, Y),R,δ,σ)
Input: S, K the explored sub-graph, Y the set of candidates vertices
Output: R the set of result patterns under construction
if $UB_2(S, K, Y) < \delta$ **then**
  return False
if $UB_3(S, K, Y) < \delta$ **then**
  return False
existsPattern $\leftarrow$ False
while $|K \cup Y| \geq \sigma$ *and* $Y \neq \emptyset$ **do**
  $v^\star = \text{argmax}_{u \in Y} A(S, K \cup \{u\})$
  $Y \leftarrow Y \setminus \{v^*\}$
  if SUB-CC*(S,* $(K \cup \{v^\star\}, Y)$*, R,*δ,σ*)* **then**
    existsPattern $\leftarrow$ True
  else
    $Y \leftarrow \{v \in Y \mid A(S, v) > A(S, v^\star)\}$
if $WRAcc(S, K) \geq \delta$ ***and*** $|K| \geq \sigma$ **then**
  existsPattern $\leftarrow$ True
  if G_K *is connected* **then**
    $R \leftarrow R \cup \{(S, K)\}$
return existsPattern

Hence, the problem we consider in this section is how to compute $\mathcal{E}'(G) \subseteq \mathcal{E}(G)$ such that $\forall (S, K) \in \mathcal{E}(G)$, $\exists (S, K') \in \mathcal{E}'(G)$ with $cov(K, K') \geq mincov$ and $\mathcal{E}'(G)$ as small as possible.

The problem of finding the set $\mathcal{E}'(G)$ of minimum size is NP hard. Thus, we adapt an efficient heuristic approach that consists in constructing $\mathcal{E}'(G)$ during the enumeration process and to use the coverage measure to prune large parts of the sub-graph search space. We use the following lower bound to stop the enumeration process of a candidate (K, Y), as it guarantees that all the patterns generated from it are covered by another already found pattern.

Definition 6 (LB)**.** *Let K be the current enumerated set of vertices, Y the set of candidate vertices, R a set of solutions ($R \subseteq \mathcal{E}(G)$), and $(S, H) \in R$. Let n and $\overline{cov}$ be:*

$$n = \lceil \delta \times \frac{|V|}{A(S, K)} \rceil - |K|$$

$$\overline{cov}(K, Y, H) = \frac{|K \cap H| + \max(0, n - |Y \setminus H|)}{|K| + \max(n, |Y \setminus H|)}$$

The lower bound function LB is thus defined by:

$$LB(K, Y, R) = \max_{(S,H) \in R} (\overline{cov}(K, Y, H))$$

This lower bound is used to prune the exploration of $\Gamma(K, Y)$ as explained below.

Property 5. *For each candidate $(K', Y') \in \Gamma(K, Y)$ such that $WRAcc(S, K') \geq \delta$, there exists at least one pattern $(S, H) \in R$ that satisfies $cov(K', H) \geq LB(K, Y, R)$.*

Proof. (Sketch) Considering that $A(S, K') \leq A(S, K)$, we can derive a bound on the minimal number of vertices that have to be added to K to obtain $WRAcc(S, K') \geq \delta$. This is the value n. If $n \leq |Y \setminus H|$, we can prove that cov is lower bounded by $\frac{|K \cap H|}{|K| + |Y \setminus H|}$. If $n > |Y \setminus H|$, there will be necessarily vertices in the intersection and cov is lower bounded by $\frac{\lceil K \cap H| + n - |Y \setminus H|}{|K| + n}$. □

Consequently, the exploration of a sub-tree $\Gamma(K, Y)$ is stopped if $LB(K, Y, R) \geq mincov$, because all the candidates (K', Y') of $\Gamma(K, Y)$ with $WRAcc(S, K') \geq \delta$ are covered by an already found solution.

C. The exceptional sub-graph space sampling approach

In practice, the end-user wants to obtain high quality patterns in a short amount of time, especially in interactive data mining processes. This can be achieved by using the solution presented in the previous section. However, this approach does not scale very well with the cardinality of the category set. To overcome this issue, we propose an approach that computes a sampling of the exceptional sub-graphs that respects the distribution of the WRAcc measure within a user-given time-budget.

We adapt the randomized pattern mining technique of [2] to exceptional sub-graphs discovery. This so called *Controlled Direct Pattern Sampling* enables the user to specify a time budget and computes a set of high quality patterns whose size directly depends on the specified amount of time. The idea consists of sampling the patterns based on a probability distribution that rewards high quality patterns. In a first attempt, we proposed to first sample the characteristics and then derive the associated sub-graphs. But this strategy failed in computing patterns with high WRAcc values because the graph structure was neglicted. Thus, we adopted the reverse approach that consists in randomly generating connected sub-graphs and then deriving the most relevant characteristic $S^\star(K)$ that fulfills the constraint $WRAcc(S^\star(K), K) \geq \delta$. We choose $S^\star(K)$ as the characteristic that contains all the characteristics S valid for K: $\forall S$ such that $valid(S, K)$, $S \subseteq S^\star(K)$. We can easily demonstrate that $S^\star(K)$ is the most relevant characteristic for K, that is $\forall S$ such that $valid(S, K)$, $WRAcc(S^\star(K), K) \geq WRAcc(S, K)$.

The sub-graph generation process is based on a random walk on a graph whose vertices are sub-graphs of G and edges (transitions) are chosen following a probability measure that favors high quality patterns $(S^\star(K), K)$. The random walk starts from an empty set ($K = \emptyset$) that is expanded by adding vertices in the neighborhood of K drawn randomly: (1) The random walk starts by drawing a first vertex. Considering the weight distribution $w_1(v) = A(S^\star(v), v)$, the probability distribution is computed by $\mathcal{P}(\{v\}) = \frac{w_1(\{v\})}{\sum_{u \in V} w_1(\{u\})}$. Thus, the more contrasted is the vertex, the greater is its probability. After drawing this first vertex v, the current sub-graph is $K = \{v\}$. (2) The random walk continues by considering the vertices in the neighborhood of the current sub-graph $G(K)$. Let $N(K)$ be the set of neighbors of K: $N(K) = \{v \in V \setminus K \mid \exists v' \in K, (v, v') \in E\}$. Then, the sub-graph generation is either stopped, or a vertex $v \in N(K)$ is drawn and added to K.

Algorithm 3: EXPRESS

Output: K the generated sub-graph
begin
  for $v \in V$ **do**
    $w_1(v) \leftarrow WRAcc(S^\star(v), v)$
  draw $v \sim w_1(v)$
  $K \leftarrow \{v\}$
  $continue \leftarrow$ True
  while $continue$ **do**
    // Calculate the probability of drawing $K \cup \{v\}$ for
    // each neighbor v:
    for $v \in N(K)$ **do**
      $K' \leftarrow K \cup \{v\}$
      $w_2(K', K) \leftarrow$
       $WRAcc(S^\star(K'), K') - WRAcc^\circ(K)$
    // Calculate the probability of drawing K:
    $w_2(K, K) \leftarrow WRAcc(S^\star(K), K) - WRAcc^\circ(K)$
    draw $K' \sim w_2(K', K)$
    if $K' = K$ **then**
      $continue \leftarrow$ False
    else
      $K \leftarrow K'$
  return K

The choice of the action is taken randomly using a weighted distribution that favors the vertex (and possibly none) whose addition to K leads to the largest increase of the WRAcc measure. Let $K' = K \cup \{v\}$. We define $WRAcc^\circ(K)$ as the lowest score that can be achieved by the current possible actions:

$$WRAcc^\circ(K) = \min(\min_{v \in N(K)} WRAcc(S^\star(K'), K'), WRAcc(S^\star(K), K))$$

From the weight $w_2(K', K)$, defined as $w_2(K', K) = WRAcc(S^\star(K'), K') - WRAcc^\circ(K)$, we derive the probability to reach the connected sub-graph $G(K \cup \{v\})$ from $G(K)$ by $\mathcal{P}(K \cup \{v\}) = \frac{w_2(K \cup \{v\}, K)}{\sum_{u \in N(K)} w_2(K \cup \{u\}, K) + w_2(K, K)}$. This distribution of probabilities rewards transitions towards connected sub-graphs K' with large $WRAcc(S^\star(K'), K')$ value. The algorithm EXPRESS[2] is repetitively called until the specified execution time is consumed. For each generated sub-graph K, if $WRAcc(S^\star(K), K) \geq \delta$, then the pattern $(S^\star(K), K)$ is added to the output result set.

IV. EXPERIMENTS

In this section, we report on experimental results to illustrate the interest of the proposed approach. We start by describing the different real-world datasets we use, as well as the questions we aim to answer. Then, we provide a performance study and give some qualitative results. The implementation of the method is in Java and the experiments run on machines equipped with i7-2600 CPUs @ 3.40GHz, and 16GB main memory, running Ubuntu 12.04, and Java Version 1.6. The code and the data are available[3].

[2]EXPRESS stands for EXcePtionnal Subgraph Sampler.

[3]https://github.com/AnesBendimerad/Exceptional-Sub-graph-Mining

A. Datasets and aims

We considered 10 real-world datasets whose characteristics are given in Table I. Eight of them come from [5] and depict Foursquare venues over 4 US and 4 EU important cities. The venues are described by a hierarchy[4]. We consider the first level (10 attributes) in the first series of experiments and the second level (around 300 attributes) for the second ones. *SF. Crimes* data[5] are provided by a Kaggle challenge and describe the criminal activity in San Francisco. Finally, *San Francisco C&V* is the combination – after normalization – of *SF. Crimes* and Foursquare data over San Francisco. Each city is divided into rectangular zones in such a way that each rectangle contains a minimal number of venues.

dataset	$\|V\|$	$\|E\|$	$\|C\|$	#objects
New York	292	647	10 (356)	71954 venues
Los Angeles	159	348	10 (325)	34504 venues
San Francisco	124	256	10 (328)	21654 venues
Washington	106	216	10 (316)	19190 venues
London	118	241	10 (318)	25029 venues
Paris	115	231	10 (305)	27443 venues
Rome	90	177	10 (279)	13166 venues
Barcelona	109	218	10 (304)	19668 venues
S.F. Crimes	898	2172	39	878049 crimes
S.F. C&V	342	767	49 (328)	878049 cr. + 21654 ven.

TABLE I: Description of the real-world datasets

ENERGETICS and EXPRESS are evaluated regarding the following questions: *What is the efficiency of* ENERGETICS *with regard to the graph characteristics that may affect its execution time? How effective are* ENERGETICS*' pruning properties? Does* ENERGETICS *scale? Does* EXPRESS *provide a good sample of Exceptional sub-graphs? What about the relevancy of Exceptional sub-graphs?*

B. Quantitative study

First, we devised a synthetic generator to evaluate how effective ENERGETICS' upper bounds are, while varying the attributed graph properties $|V|, |E|$ and $|C|$. A baseline algorithm is obtained by deactivating the upper bounds and the pruning abilities of ENERGETICS. Thus, the baseline only pushes monotonic constraints. We also ran ENERGETICS using each upper-bound in turn. Results are given in Fig. 3. As the baseline only works with very small graphs, we report on ENERGETICS' behavior with much more larger characteristics in the last column. In most of the cases, ENERGETICS algorithm outperforms the baseline and the versions with a single upper-bound by several orders of magnitude. The use of a single upper-bound takes in some cases more time than the baseline computation. This is due to the upper-bound verification that can be expensive while ineffective if not used in conjunction to other upper-bounds. There is no upper bound that outperforms the others. The efficiency of our algorithm is due to the conjunctive use of all the upper-bounds. We can observe that ENERGETICS scales well on the synthetic data according to the number of vertices and the number of edges while its execution time increases exponentially w.r.t. to the number of attributes.

[4]https://developer.foursquare.com/categorytree

[5]https://www.kaggle.com/c/sf-crime

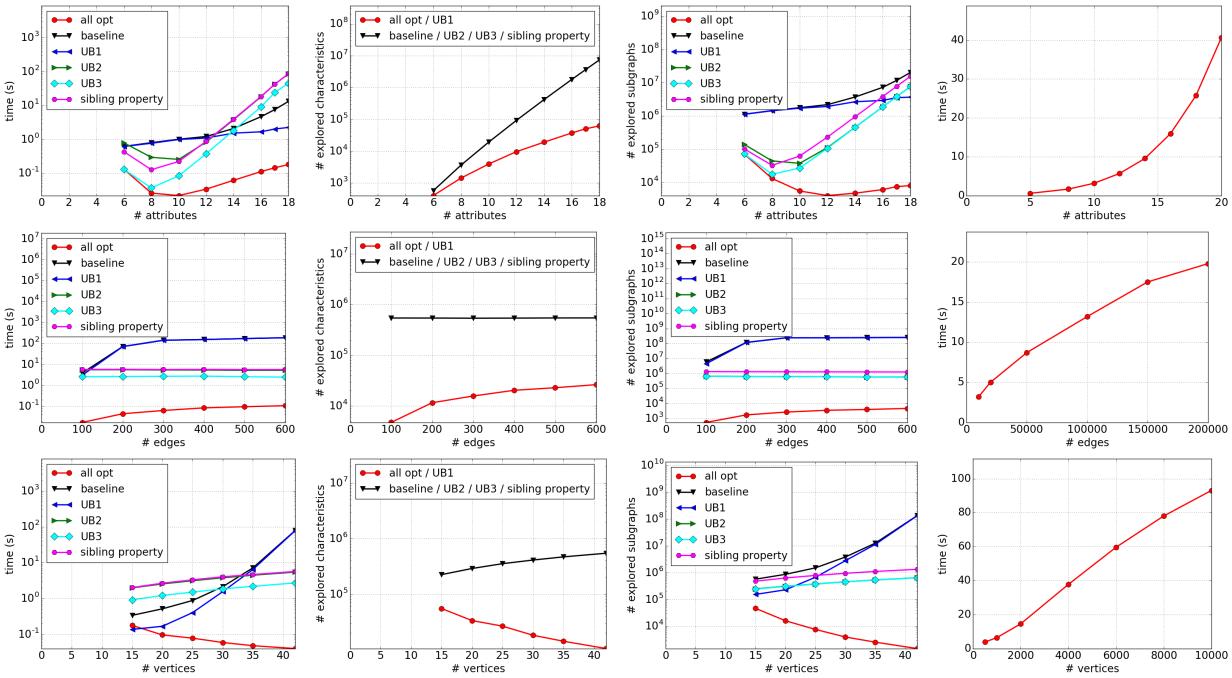


Fig. 3: Impact of the upper bounds on runtime (1st column), search space size – characteristics (2nd column), and subgraph (3rd column) w.r.t. the number of attributes (1st row), edges (2nd row), and vertices (3rd row). Broader parameter values are explored for ENERGETICS in the 4th column. Experiments carried out on synthetic data with $\delta = 0.015, \sigma = 1, mincov = 0.8$.

Fig. 4 reports the behavior (runtime, number of patterns, and number of explored subgraphs) of ENERGETICS on real-world datasets when varying the input parameters ($mincov$, δ and σ). Notice that ENERGETICS fails on the real-world data when the constraint on $mincov$ is desactivated. We can observe that the behavior of ENERGETICS remains unchanged when varying $mincov$. This is due to the fact that most of redundant subgraphs are strictly included in other ones and they are discarded with $mincov = 1$. Results for the two other parameters δ and σ are as expected: The execution time increases when either the thresholds σ and δ become less stringent. *S.F. C&V* is the dataset whose execution times are the most important. It confirms the fact that the number of attributes is the most influential data parameter.

We also studied the behavior of our algorithm w.r.t. the replication factor. For a replication factor equal to n, the attributed graphs are duplicated n times, so the initial vertices are repeated n times with the same attributes values and the same connections with the corresponding duplicated vertices. Fig. 5 presents the ratio of the execution time divided by the execution time on the original dataset. Thus, the ratio is equal to 1 for $n = 1$. We can observe that ENERGETICS performances does not degrade too much when the dataset size increases.

These first experiments demonstrate that ENERGETICS is only efficient for graphs whose number of attributes is small. EXPRESS has been designed especially to perform on graphs with hundreds of attributes, using a time budget to control the execution time and the number of computed patterns. To evaluate the ability of EXPRESS to compute exceptional sub-graphs of high $WRAcc$ values, we report on Fig. 6 the distributions of the WRAcc measure on both the complete set of exceptional sub-graphs returned by ENERGETICS and the sample set provided by EXPRESS. Several time budgets are used and are all much lower than the execution time required by ENERGETICS. We can observe that the two distributions are similar and thus the sampling approach succeeds in fostering patterns with high WRAcc measure. Also, the higher the time budget, the better the distribution. Fig. 7 reports similar distributions for the real-world datasets with hundreds number of attributes for which an exhaustive search is not possible. The distributions are similar. Thus, EXPRESS makes it possible to discover high quality patterns within a time-budget.

C. Qualitative study

ENERGETICS was applied on Barcelona graph with 10 attributes. Fig. 8 (left) displays 5 patterns discovered. Pattern P_1 depicts neighborhoods with a high concentration of venues of type *Outdoors & Recreation*. Most of these zones are near the sea with the Olympic harbor and the main beaches of Barcelona. This pattern overlaps with two other patterns, P_3 and P_4. Notice that the ICDM'16 conference venue is covered by both P_1 and P_3. The attendees can take benefits of the high concentration of *Outdoors & Recreation* places as well as the *nightlife spots* and the *food* places. However, the concentration of *shops* in this areas is lower than in the rest of the city. Pattern P_2 depicts zones with a high concentration of places

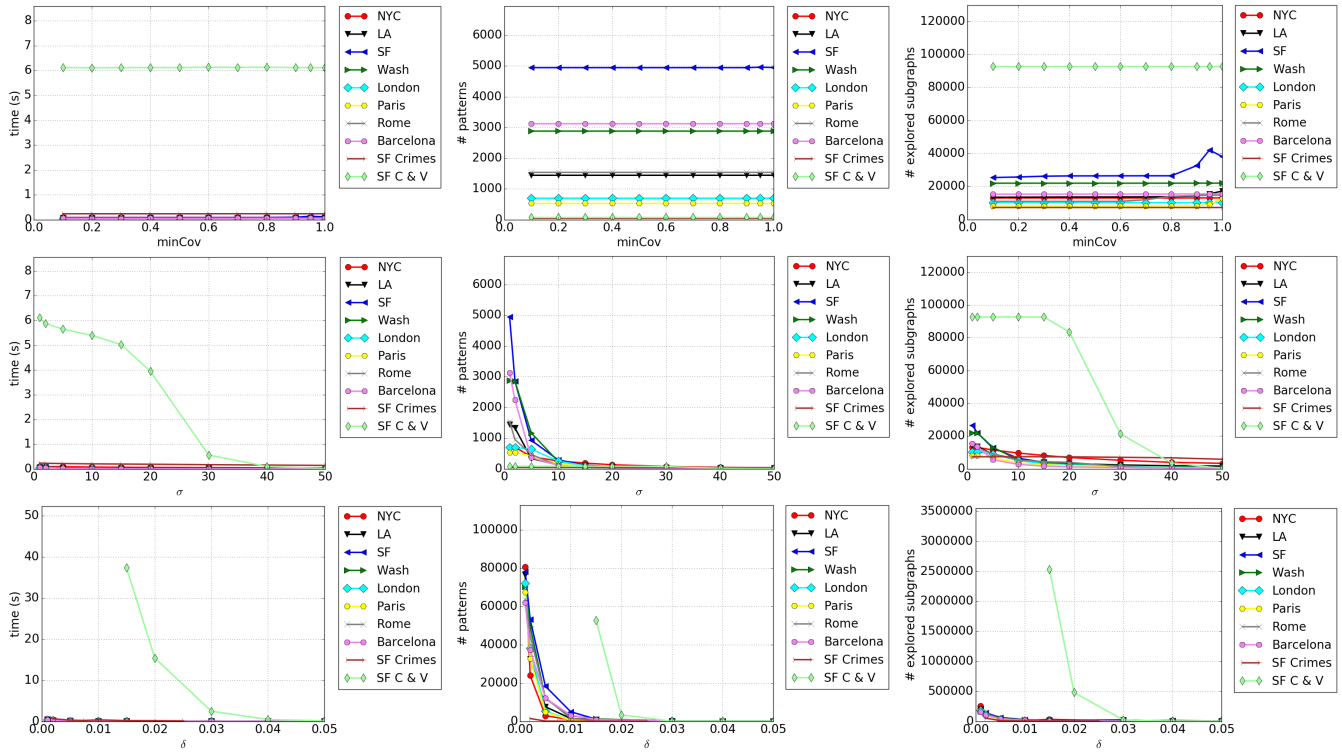


Fig. 4: Behavior of ENERGETICS (runtime in 1st column, #patterns in 2nd column and, #explored subgraphs in 3rd column) according to $mincov$ (1st row), σ (2nd row) and δ (3rd row) for the 10 real-world datasets ($\delta = 0.001, \sigma = 1, mincov = 0.8$).

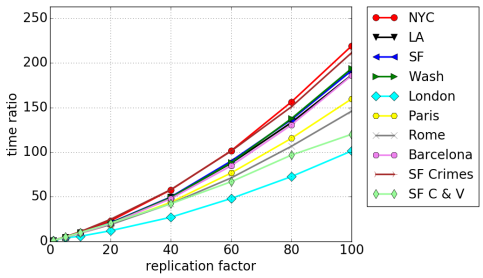


Fig. 5: Runtime ratio w.r.t. the replication factor for real world datasets ($\delta = 0.001$ except SF Crimes and SF C&V (0.007 and 0.03), $\sigma = 1, mincov = 0.8$.

of the type *Outdoors & Recreation* or *College & Universities* while the concentration of *food* places is low. This pattern is sensible since it contains the famous *camp nou* stadium as well as several universities and colleges.

We also applied EXPRESS on Barcelona graph with 304 attributes. Some patterns are reported in Fig. 8 (right). Pattern P_6 identifies an area with a high concentration of *metro stations* and *state and municipalities* as well as *Home (private)* in the North of the city. P_7 depicts an area with a high concentration of *bars* and typical *Mediterranean restaurants* while P_8 highlights areas with a larger variety of restaurants (e.g., *Salad Place, Asian and Italian Restaurants*). P_9 contains a high concentration of *hotels and bars*. Notice that P_7, P_8 and P_9 are closed to ICDM'16 conference venue which attendees could take advantage of these patterns to prepare their venue according to their preferences.

Besides, we mined exceptional sub-graphs on the different cities. In most of them (e.g., Barcelona, Paris, Rome, Los Angeles, London), the *nightlife spots* are mainly located in the city center. For New York, they are in the south of Manhattan and the west of Brooklyn. The higher concentration of *outdoor & recreation* places is surrounding for London. For seaside towns, they are concentrated on the coasts.

V. RELATED WORK

Several approaches have been designed to discover new insights in vertex attributed graphs. The pioneering work of Moser et al. [14] presents a method to mine dense homogeneous sub-graphs, i.e., sub-graphs whose vertices share a large set of attributes. Similar to that work, Günnemann et al. [7] introduce a method based on subspace clustering and dense sub-graph mining to extract non redundant sub-graphs that are homogeneous with respect to the vertex attributes. Silva et al. [18] extract pairs made of a dense sub-graph and a Boolean attribute set such that the Boolean attributes are strongly associated with the dense sub-graphs. In [16], the authors propose to mine the graph topology of a large attributed graph by finding regularities among numerical vertex descriptors. The main objective of all these approaches is to find regularities instead of peculiarities within a large graph,

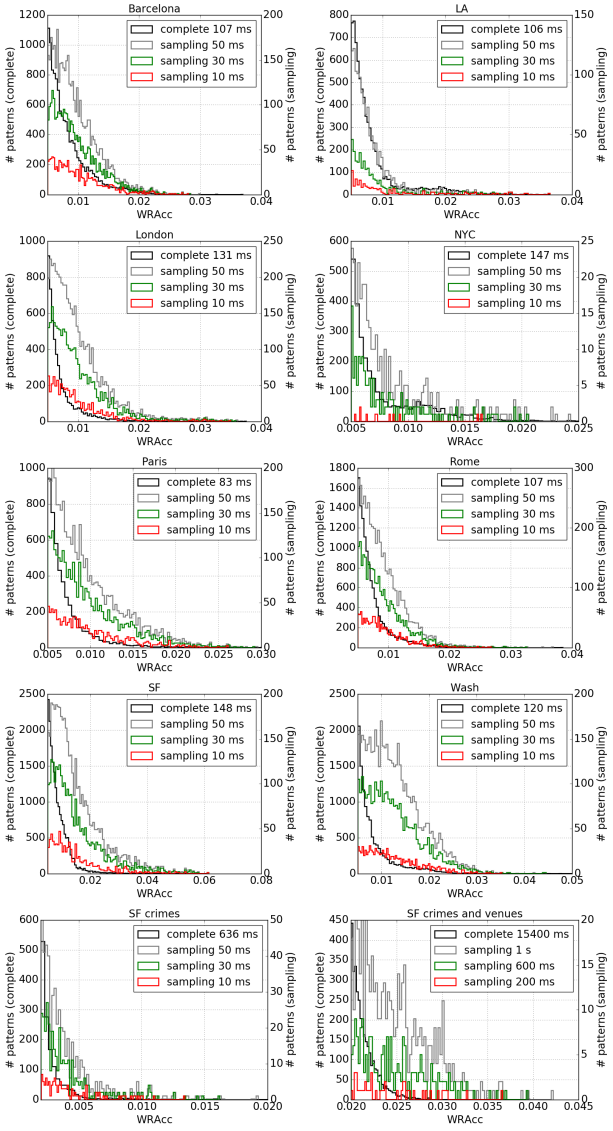


Fig. 6: Distributions of the patterns from ENERGETICS and EXPRESS with different time budgets ($\delta = 0.001$).

whereas *Exceptional Sub-graph Mining* computes sub-graphs with their distinguishing characteristics.

Interestingly, a recent work [1] proposes to mine descriptions of communities from vertex attributes, with a Subgroup Discovery approach. In this supervised setting, each community is treated as a target that can be assessed by well-established measures, as the WRAcc measure used in this paper. More generally, Subgroup Discovery [9], [15] aims to find descriptions of subpopulations for which the distribution of a predefined target value is significantly different from the distribution in the whole data. When there are multiple targets, Subgroup Discovery consists in finding descriptions that significantly change the joint distribution of the target attributes – a task that has been introduced as *Exceptional Model Mining* [10], [3]. A variety of measures of changes

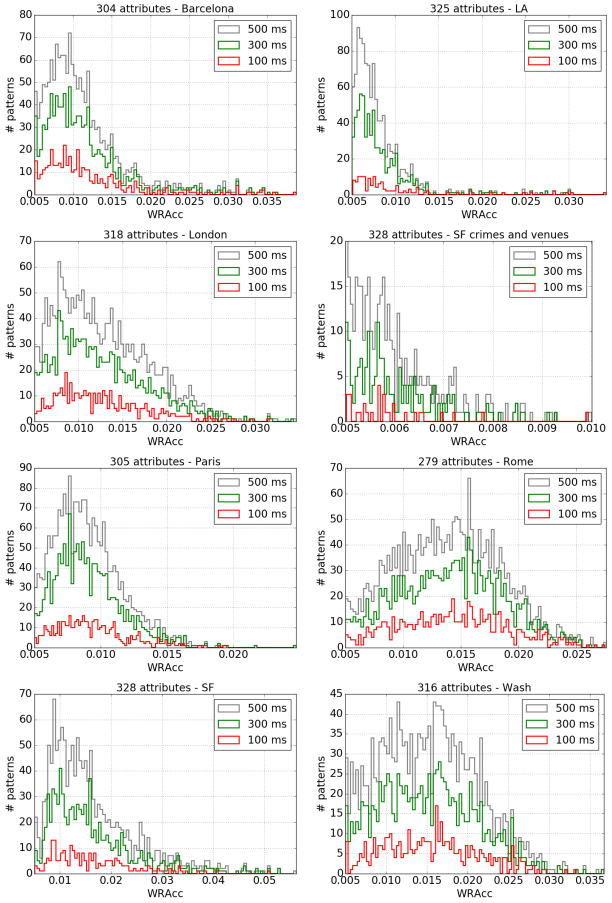
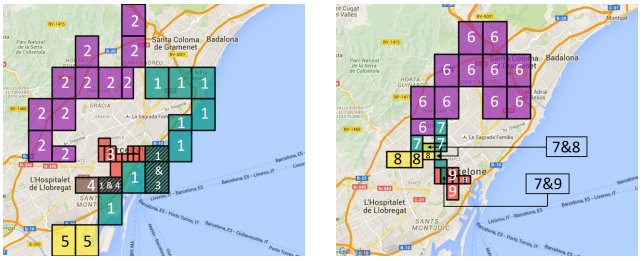


Fig. 7: Distribution of the output space sampling with different time budgets for datasets with larger number of attributes.

have been used: pairwise correlation and entropy measures [10], Bayesian networks [4], Kullback-Leibler Divergence and encoding difference based on Minimum Description Length [21]. The common point to all these approaches is that the combination of large target space and non-monotonic measures leads to the use of heuristic search methods, i.e., beam search. Furthermore, these methods are supervised since the target attributes are given. The algorithms proposed in this paper extend many of these results to the unsupervised setting.

Motivated by both scalability issue and user interaction needs, sampling the output space of patterns has received much attention in the past decade. Many approaches have been proposed in the context of transactional data [2], [13], [12], [6], [11]. Some researchers have tackled the problem of sampling the output space of frequent sub-graphs in a collection of graphs [8], [17]. These methods are based on random walks. In particular, [17] aims at returning the top k frequent graphs of a specified size. The problem we tackle is different on several points: We consider a single graph and a discriminative measure instead of a frequency measure. Beside, these methods aim at sampling frequent sub-graphs while we address the problem of exceptional attributed sub-



Id	S^+	S^-	$WRAcc$	size
P_1	Outdoors & Recreation	Shop & Service, Professional & Other Places	0.0186	11
P_2	Outdoors & Recreation, College & University	Food	0.0223	11
P_3	Nightlife Spot, Food	Professional & Other Places	0.0268	14
P_4	Outdoors & Recreation, Events, Art & Entertainment	Shop & Service, College & University, Travel & Transport, Food	0.0158	2
P_5	Shop & Service, Professional & Other Places	Outdoors & Recreation, Event, Nightlife Spot, Art & Entertainment, Food	0.0179	2
P_6	Metro Station, Home (private), States & Municipalities		0.0056	11
P_7	Mediterranean Restaurant, Bar		0.0055	9
P_8	Salad Place, Office, Asian Restaurant, Italian Restaurant		0.0043	5
P_9	Hotel, Bar		0.006	8

Fig. 8: Patterns discovered in Barcelona datasets: by ENERGETICS with 10 attributes (patterns P_1 to P_5 plotted on the left-hand side map) and by EXPRESS with 304 attributes (patterns P_6 to P_9 plotted on the right-hand side map).

graph sampling which is much more challenging since we have to deal simultaneously with two dimensions: sub-graphs and characteristics. Our approach is based on a random walk over sub-graphs that fosters patterns with a high WRAcc measure.

VI. CONCLUSION

We introduced the exceptional sub-graph mining problem to discover homogenous areas that differ from the rest of the city. We defined an efficient algorithm that computes the complete set of exceptional sub-graphs by taking advantage of several tight upper bounds and other pruning properties. We also introduced an additional constraint to avoid redundancy. Eventually, we designed an algorithm to sample the output space of patterns to enable time-budget analysis. We reported a extensive empirical study over 10 real-world datasets that demonstrates the relevancy of our proposal. Whereas we demonstrate the efficiency of the pruning techniques, ENERGETICS still has difficulties to scale with the number of attributes. This problem is fixed by EXPRESS that has capabilities to mine graphs described by hundreds of attributes while preserving the WRAcc distribution. We also illustrated the relevancy of the discovered patterns thanks to a qualitative analysis. Our proposal can be extended to take into account other graph topological properties (e.g., diameter). Other quality measure can be investigated to assess the quality of the characteristics in the city areas and highlight some specific phenomena.

ACKNOWLEDGMENT

This work was supported in part by the Group Image Mining (GIM) which joins researchers of THALES Group and LIRIS Lab. We thank especially Jérôme Kodjabachian and Bertrand Duqueroie of AS&BSIM Lab. of THALES Group. This work is also partially supported by the EU FP7-PEOPLE-2013-IAPP project GRAISearch and the CNRS project PEPS-2015-Préfute.

REFERENCES

[1] Martin Atzmueller, Stephan Doerfel, and Folke Mitzlaff. Description-oriented community detection using exhaustive subgroup discovery. *Inf. Sci.*, 329:965–984, 2016.
[2] Mario Boley, Claudio Lucchese, Daniel Paurat, and Thomas Gärtner. Direct local pattern sampling by efficient two-step random procedures. In *ACM SIGKDD 2011*, pages 582–590, 2011.
[3] Wouter Duivesteijn, Ad Feelders, and Arno J. Knobbe. Exceptional model mining - supervised descriptive local pattern mining with complex target concepts. *Data Min. Knowl. Discov.*, 30(1):47–98, 2016.
[4] Wouter Duivesteijn, Arno J. Knobbe, Ad Feelders, and Matthijs van Leeuwen. Subgroup discovery meets bayesian networks – an exceptional model mining approach. In *ICDM 2010*, pages 158–167, 2010.
[5] Géraud Le Falher, Aristides Gionis, and Michael Mathioudakis. Where is the soho of Rome? Measures and algorithms for finding similar neighborhoods in cities. In *ICWSM 2015*, pages 228–237, 2015.
[6] Arnaud Giacometti and Arnaud Soulet. Frequent pattern outlier detection without exhaustive mining. In *PAKDD 2016*, pages 196–207, 2016.
[7] Stephan Günnemann, Ines Färber, Brigitte Boden, and Thomas Seidl. Subspace clustering meets dense subgraph mining. In *ICDM 2010*, pages 845–850, 2010.
[8] Mohammad Al Hasan and Mohammed J. Zaki. Output space sampling for graph patterns. *PVLDB*, 2(1):730–741, 2009.
[9] Nada Lavrac, Branko Kavsek, Peter A. Flach, and Ljupco Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.
[10] Dennis Leman, Ad Feelders, and Arno J. Knobbe. Exceptional model mining. In *ECML/PKDD 2008*, pages 1–16, 2008.
[11] Geng Li and Mohammed J. Zaki. Sampling frequent and minimal boolean patterns. *Data Min. Knowl. Discov.*, 30(1):181–225, 2016.
[12] Sandy Moens and Mario Boley. Instant exceptional model mining using weighted controlled pattern sampling. In *IDA*, pages 203–214, 2014.
[13] Sandy Moens and Bart Goethals. Randomly sampling maximal itemsets. In *SIGKDD Workshop on Interactive Data Exploration and Analytics*, pages 79–86. ACM, 2013.
[14] Flavia Moser, Recep Colak, Arash Rafiey, and Martin Ester. Mining cohesive patterns from graphs with feature vectors. In *SDM 2009*, pages 593–604, 2009.
[15] Petra Kralj Novak, Nada Lavrac, and Geoffrey I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *J. Mach. Learn. Res.*, 10:377–403, 2009.
[16] Adriana Prado, Marc Plantevit, Céline Robardet, and Jean-François Boulicaut. Mining graph topological patterns: Finding covariations among vertex descriptors. *IEEE TKDE.*, 25(9):2090–2104, 2013.
[17] Tanay K. Saha and Mohammad A. Hasan. A sampling based method for top-*k* frequent subgraph mining. *Stat. An. & DM*, 8(4):245–261, 2015.
[18] Arlei Silva, Wagner Meira Jr., and Mohammed J. Zaki. Mining attribute-structure correlated patterns in large attributed graphs. *PVLDB*, 5(5):466–477, 2012.
[19] Seth E. Spielman and Jean-Claude Thill. Social area analysis, data mining, and GIS. *Comp. Env. & Urb. Sys.*, 32(2):110–122, 2008.
[20] Takeaki Uno. An efficient algorithm for enumerating pseudo cliques. In *ISAAC 2007*, pages 402–414, 2007.
[21] Matthijs van Leeuwen. Maximal exceptions with minimal descriptions. *Data Min. Knowl. Discov.*, 21(2):259–276, 2010.
[22] Jia Wang, James Cheng, and Ada Wai-Chee Fu. Redundancy-aware maximal cliques. In *ACM SIGKDD 2013*, pages 122–130, 2013.