Mining exceptional closed patterns in attributed graphs

Anes Bendimerad, Marc Plantevit, Céline Robardet

Noname manuscript No.
(will be inserted by the editor)

Mining Exceptional Closed Patterns in Attributed Graphs

Anes Bendimerad · Marc Plantevit · Céline Robardet

Received: date / Accepted: date

Abstract Geo-located social media provide a large amount of information describing 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 suitably attributed graph model, our approach identifies neighborhoods with homogeneous and exceptional characteristics. We introduce the novel problem of exceptional subgraph mining in attributed graphs and propose a complete algorithm that takes benefits from closure operators, new upper bounds and pruning properties. We also define an approach to sample the space of closed exceptional subgraphs within a given time-budget. Experiments performed on 10 real datasets are reported and demonstrate the relevancy of both approaches, and also show their limits.

Keywords Exceptional subgraph mining, pattern mining, urban data analysis.

Anes Bendimerad
INSA Lyon, CNRS
LIRIS UMR5205
F-69621 France
E-mail: ahmed-anes.bendimerad@liris.cnrs.fr

Marc Plantevit
Université Lyon 1, CNRS
LIRIS UMR5205
F-69622 France
E-mail: marc.plantevit@liris.cnrs.fr

Céline Robardet
INSA Lyon, CNRS
LIRIS UMR5205
F-69621 France
E-mail: celine.robardet@liris.cnrs.fr

1 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 [8], matching social attributes with geographic spaces [30] or characterization of neighborhoods for analyzing urban mobility [25].

Using social and urban data of a city (such as the ones provided by social networks as FOURSQUARE or GOOGLEPLACE), 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 subgraphs 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, [22] designed a method to find dense homogeneous subgraphs, where vertices are described by categorical attributes and [10] proposes subspace clustering approach using numerical vertex attributes. However, all these works focus their attention on the similarity inside the subgraphs, while underestimating exceptionality of the subgraph characteristics with respect to the whole graph.

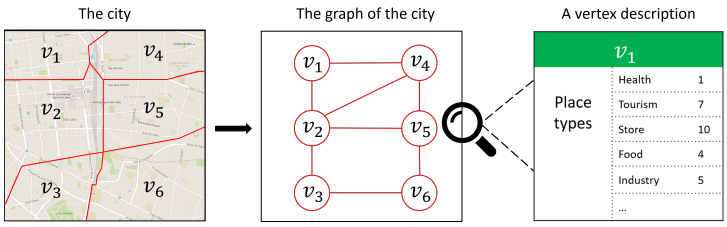


Fig. 1: Example of a graph modeling a city.

1.1 Research contribution

We propose two algorithms to discover exceptional subgraphs. The first one is an exact algorithm that uses original and efficient upper bounds and some other techniques to reduce the search space. It takes benefit from a closure operator to avoid redundancy and efficiently prune the search space. The second algorithm mines closed exceptional subgraphs by directly sampling the space of closed patterns in a similar way as [3, 11, 28].

Our main contributions are manifold:

- 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.
- We present an efficient algorithm based on new upper bounds and pruning properties to discover exceptional subgraphs.
- We design a probabilistic approach that directly samples the output space of patterns within a time budget specified by end-users.
- 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 relevance of the discovered results.

This paper extends our previous work [2]. First, a slight modification of the quality measure enables to discover more relevant patterns. Second, the introduction of a closure operator makes it possible to define a more efficient algorithm by removing redundant parts in the search space. Experiments give evidence that our novel algorithm outperforms the algorithm defined in [2] with several orders of magnitude.

1.2 Outline

This paper is organized as follows. The first section formally introduces the problem. The proposed solutions are presented in Section 3. We report a systematic empirical study on numerous real-world datasets in Section 4. Section 5 discusses related work and our conclusions are drawn in the final section.

2 Problem Setting

In this section, we provide the necessary definitions and terminology. Table 1 summarizes the definitions of the symbols used in the paper. 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),

Symbols	Definitions
$G = (V, E, C, D)$	An attributed graph with vertex set V and edge set E. The vertices in V are described by count variables whose labels are denoted $C = \{c_1 \dots, c_p\}$ and values $D = \{c_1(v), \dots, c_p(v) \mid v \in V\}$.
L	A set of labels: $L \subseteq C$.
K	A set of vertices: $K \subseteq V$.
$S = (S^+, S^-)$	A characteristic: $S^+, S^- \subseteq C$, $S^+ \cap S^- = \emptyset$
$\mathcal{S}$	Then set of all characteristics on C.
$G[K]$	The subgraph of G induced by $K \subseteq V$.
f and g	The Galois connection between 2^V and $\mathcal{S}$.

Table 1: Symbol table.

$C = \{c_i,\ i \in [\![1, p]\!]\}$ 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. $sum(health, \{v_1\}) = 1$ 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 $K = \{v_2, v_5\}$, $sum(K) = 49$.

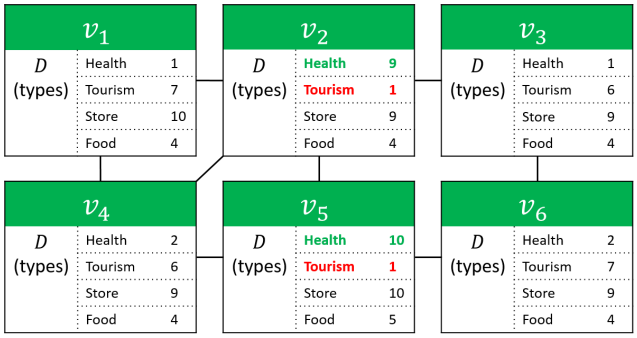


Fig. 2: Example of the distribution of venues in areas.

Our objective is to identify neighborhoods whose characteristics distinguish them from the rest of the city. To that end, we propose to discover connected

subgraphs associated to exceptional categories. A category is exceptional for a subgraph 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 element to describe a neighborhood. For example, 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*. We formalize the excess and deficit in the amount of some categories by means of characteristics defined as

Definition 1 (Characteristic) A characteristic is defined as a pair $S = (S^+, S^-)$ with S^+ and S^- two disjoint subsets of C. The set of all characteristics is denoted $\mathcal{S}$. We also define operators between two characteristics $S_1 = (S_1^+, S_1^-)$ and $S_2 = (S_2^+, S_2^-)$:

- $S_1 \cap S_2 = (S_1^+ \cap S_2^+, S_1^- \cap S_2^-)$
- $S_1 \cup S_2 = (S_1^+ \cup S_2^+, S_1^- \cup S_2^-)$
- $S_1 \subseteq S_2 \Leftrightarrow S_1^+ \subseteq S_2^+ \wedge S_1^- \subseteq S_2^-$
- $|S| = |S^+| + |S^-|$

In order to assess the relevancy of the characteristic S with respect to the subgraph 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 [14].

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 a 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 [14], 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 subgraph.

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

The main differences with the $WRAcc$ used in Subgroup Discovery [14] are (1) our adapted $WRAcc$ considers both the positive and the negative contrasts

in an unsupervised setting (i.e., there is no class attribute in our setting, the "target" is settled by each pattern), (2) it takes into account the homogeneity of elements of K, using the predicate $valid(S, K)$.

In [2], we used a slightly different $Wracc$ measure that differs by its normalization factor (i.e., $\frac{|K|}{|V|}$ was used instead of $\frac{sum(K)}{sum(V)}$ in this paper). This new coefficient makes it possible to correct the defect of the previous measure consisting in fostering sparse areas.

We now define the pattern domain we consider:

Definition 2 (Exceptional subgraph) Given a graph $G = (V, E, C, D)$ and two thresholds σ and δ, an exceptional subgraph (S, K) is such that (1) $|K| \geq \sigma$, (2) $G[K]$ is connected, and (3) $WRAcc(S, K) \geq \delta$.

Given an exceptional subgraph (S, K), a large number of less specific subgraphs can be derived, i.e. patterns (S', K') such that $S' \subseteq S$ and $K' \subseteq K$. As these patterns (S', K') are already described and covered by (S, K), they unnecessarily increase the size of the solution set. This redundancy can be avoided thanks to a closure operator [13] defined below.

Definition 3 (Formal concept) Let f and g be two closure operators forming a Galois connection:

- $f : 2^V \rightarrow \mathcal{S}$, that provides the most specific characteristic associated to the subgraph induced by $K \subseteq V$:

$$f(K) = \Big(\{c_i \in C \mid \bigwedge_{v \in K} \delta_{gain(c_i,v)>0}\}, \{c_i \in C \mid \bigwedge_{v \in K} \delta_{gain(c_i,v)<0}\}\Big)$$

- $g : \mathcal{S} \rightarrow 2^V$, that returns the set of vertices supporting the characteristic S:

$$g(S) = \{v \in V \mid valid(S, \{v\})\}$$

A pair (S, K), with $S \in \mathcal{S}$ and $K \subseteq V$, is a formal concept iff $S = f(g(S))$ and $K = g(S)$, or equivalently, $S = f(K)$ and $K = g(f(K))$.

It may happen that a formal concept as defined above does not correspond to a connected subgraph. For example, in Fig. 3, (S, K) is a formal concept, with $S = (\{c_1\}^+, \{c_2\}^-)$ and $K = \{v_1, v_3, v_4, v_6\}$. However, (S, K) is not an exceptional subgraph because $G[K]$ is not connected. Maximal patterns address this limitation:

Definition 4 (Maximal pattern) A set of maximal patterns is derived from a formal concept (S, K) as:

$$\{(f(CC), CC) \mid CC \text{ is a connected component of } G[K]\}$$

In other terms, a maximal pattern $(f(CC), CC)$ is made of the most specific characteristic for CC, but also, the connected subgraph $G[CC]$ cannot be extended to another connected subgraph while keeping the current characteristic $f(CC)$.

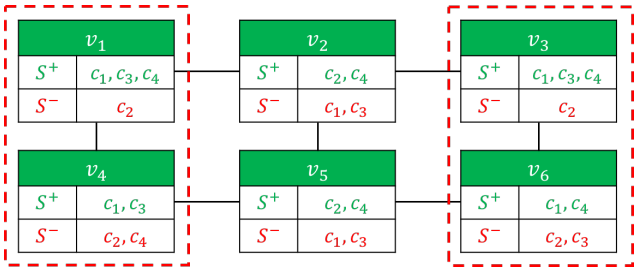


Fig. 3: Example of a formal concept (S, K), with $S = (\{c_1\}^+, \{c_2\}^-)$, $K = \{v_1, v_3, v_4, v_6\}$ such that $G[K]$ is not connected.

Following our example in Fig. 3, the formal concept (S, K) contains two connected components $CC_1 = \{v_1, v_4\}$ and $CC_2 = \{v_3, v_6\}$, with $f(CC_1) = (\{c_1, c_3\}^+, \{c_2\}^-)$ and $f(CC_2) = (\{c_1, c_4\}^+, \{c_2\}^-)$. From these two connected components, two maximal patterns $(f(CC_1), CC_1)$ and $(f(CC_2), CC_2)$ are derived.

Finally, all these definitions are used to establish the notion of closed exceptional subgraph:

Definition 5 (Closed exceptional subgraph) Let $S \in \mathcal{S}$ be a characteristic and $K \subseteq V$ a subset of vertices, (S, K) is a closed exceptional subgraph iff (1) (S, K) is a maximal pattern (2) (S, K) is an exceptional subgraph.

The rest of the paper is devoted to the computation and evaluation of the complete set of closed exceptional subgraphs. This requires searching for 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 a large number of categories. In the following, we propose an efficient approach that takes benefit from closed pattern properties.

3 Computing exceptional subgraphs

This section introduces two distinct approaches to extract closed exceptional subgraphs. First, we present an exact algorithm that aims at discovering the complete set of closed exceptional subgraphs. Second, we devise a heuristic algorithm that samples the space of closed exceptional subgraphs within a user-defined time-budget. This approach makes possible to obtain instant results and to successfully scale up on datasets with a large number of attributes.

3.1 The complete approach

In order to enumerate the set of all closed exceptional subgraphs, we explore the space of characteristics $S = (S^+, S^-)$, and for each characteristic, we enumerate the maximal patterns that can be generated from S using the closure

operators. We start from an empty characteristic $(S^+, S^-) = (\emptyset, \emptyset)$ and consider the candidate categories that can be used to expand S: $X = (X^+, X^-)$: X^+ contains the categories that can be added to S^+, and X^- the ones that can be added to S^-. $Y \subseteq V$ represents a set of vertices that verifies $valid(S, Y)$. Initially, Y contains all the vertices V, and $(X^+, X^-) = (C, C)$. In each recursive call of CEnergetics, S is extended with an element x of X^+ or of X^-. Y is then reduced to the vertices v that satisfy $valid(S \cup \{x\}, \{v\})$.

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 $K \subseteq V$, we have $valid(S_2, K) \Rightarrow valid(S_1, K)$. By the contraposition, the invalid vertices for S_1 are also invalid for S_2, and therefore, the valid set of vertices associated to $S \cup \{x\}$ is a subset of Y (Line 5). We also take benefit from this anti-monotony using the fail first principle: To extend the current characteristic S, we choose the characteristic x for which the set $\{v \in Y \mid valid(S \cup \{x\}, \{v\})\}$ is the smallest. After updating Y, we explore each connected component CC of $G[Y]$ independently and form $(f(CC), CC)$ that is, by definition, a maximal pattern. If $f(CC) \subseteq S \cup X$, then the maximal pattern $(f(CC), CC)$ has not yet been explored and CEnergetics is recursively called with $S = f(CC)$ and $Y = CC$ (Line 9). This allows to explore only characteristics S and vertices subsets Y that form maximal patterns (S, Y), and without redundancy.

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

Input: $S = (S^+, S^-)$ the current explored characteristic, $X = (X^+, X^-)$ the candidate sets, Y a connected component s.t (S, Y) is a maximal pattern
Output: R the result set under construction

```
1  if X ≠ (∅, ∅) then
2  |  if |Y| ≥ σ and UB(S ∪ X, Y) ≥ δ then
3  |  |  // Extending S using the fail first principle:
4  |  |  x ← argmin_{x∈X} |{v ∈ Y | valid(S ∪ {x}, {v})}|
5  |  |  Y' ← {v ∈ Y | valid(S ∪ {x}, {v})}
6  |  |  for each connected component CC ⊆ G[Y'] do
7  |  |  |  if f(CC) ⊆ S ∪ X then
8  |  |  |  |  // (f(CC), CC) has not been explored yet
9  |  |  |  |  CEnergetics(f(CC), X \ f(CC), CC, R, δ, σ)
10 |  |  CEnergetics(S, X \ {x}, Y, R, δ, σ)
11 else
12 |  if |Y| ≥ σ and WRAcc(S, Y) ≥ δ then
13 |  |  R ← R ∪ {(S, Y)}
```

Another pruning mechanism is used on Line 2 where the function UB is used to upper bound the $WRAcc$ measure. This function relies on the aggregation property of the $WRAcc$ measure as defined below.

Property 1 Let $S = (S^+, S^-)$ be a characteristic, and $K \subseteq V$ a set of vertices satisfying $valid(S, K)$. We have:

$$WRAcc(S, K) = \sum_{v \in K} \sum_{x \in S} WRAcc(\{x\}, \{v\})$$

Proof Since $valid(S, K) = true$:

$$
\begin{aligned}
WRAcc(S,K) = A(S,K) \times \frac{sum(K)}{sum(V)} \\
&= \left(\frac{sum(S^+,K)}{sum(K)} - \frac{sum(S^+,V)}{sum(V)} - \frac{sum(S^-,K)}{sum(K)} + \frac{sum(S^-,V)}{sum(V)}\right) \times \frac{sum(K)}{sum(V)} \\
&= \frac{sum(S^+,K) - sum(S^-,K)}{sum(V)} - \frac{sum(S^+,V) - sum(S^-,V)}{sum(V)^2} \times sum(K) \\
&= \sum_{v \in K} \left(\frac{sum(S^+,\{v\}) - sum(S^-,\{v\})}{sum(V)}\right) - \sum_{v \in K} \left(\frac{sum(S^+,V) - sum(S^-,V)}{sum(V)^2} \times sum(\{v\})\right) \\
&= \sum_{v \in K} \left(\frac{sum(S^+,\{v\}) - sum(S^-,\{v\})}{sum(V)} - \frac{sum(S^+,V) - sum(S^-,V)}{sum(V)^2} \times sum(\{v\})\right) \\
&= \sum_{v \in K} \left(\sum_{x \in S} \left(\frac{sum(x^+,\{v\}) - sum(x^-,\{v\})}{sum(V)}\right) - \sum_{x \in S} \left(\frac{sum(x^+,V) - sum(x^-,V)}{sum(V)^2} \times sum(\{v\})\right)\right) \\
&= \sum_{v \in K} \sum_{x \in S} \left(\frac{sum(x^+,\{v\}) - sum(x^-,\{v\})}{sum(V)} - \frac{sum(x^+,V) - sum(x^-,V)}{sum(V)^2} \times sum(\{v\})\right) \\
&= \sum_{v \in K} \sum_{x \in S} \left(\left(\frac{sum(x^+,\{v\})}{sum(\{v\})} - \frac{sum(x^+,V)}{sum(V)} - \frac{sum(x^-,\{v\})}{sum(\{v\})} + \frac{sum(x^-,V)}{sum(V)}\right) \times \frac{sum(\{v\})}{sum(V)}\right) \\
&= \sum_{v \in K} \sum_{x \in S} WRAcc(\{x\},\{v\})
\end{aligned}
$$

From this property, we can derive the following function UB and demonstrate that it can be used to upper bounds the $WRAcc$ value.

Definition 6 (UB) Let $S = (S^+, S^-)$ be a characteristic, and $K \subseteq V$. $UB(S, K)$ is defined as:

$$UB(S,K) = \sum_{v \in K} \sum_{x \in S} WRAcc(\{x\},\{v\})$$

Property 2 For each pattern (S_2, K_2) such that $S_2 \subseteq S$ and $K_2 \subseteq K$, we have

$$UB(S,K) \geq WRAcc(S_2,K_2)$$

Proof (1) If $valid(S_2, K_2) = false$, the property is verified because $UB(S, K) \geq 0$. In fact, UB is a sum of $WRAcc$ values that are always positive or null.

(2) If $valid(S_2, K_2) = true$:

$$\begin{aligned} UB(S,K) &= \sum_{v \in K_2} \sum_{x \in S_2} WRAcc(\{x\},\{v\}) + \sum_{v \in K \setminus K_2} \sum_{x \in S_2} WRAcc(\{x\},\{v\}) \\ &+ \sum_{v \in K} \sum_{x \in S \setminus S_2} WRAcc(\{x\},\{v\}) \\ &= WRAcc(S_2,K_2) + \sum_{v \in K \setminus K_2} \sum_{x \in S_2} WRAcc(\{x\},\{v\}) \\ &+ \sum_{v \in K} \sum_{x \in S \setminus S_2} WRAcc(\{x\},\{v\}) \\ &\geq WRAcc(S_2,K_2) \end{aligned}$$

Since all the enumerated patterns $P = (S_2, K_2)$ by CEnergetics satisfy $S_2 \subseteq S \cup X$ and $K_2 \subseteq Y$, we have always $UB(S \cup X, Y) \geq WRAcc(S_2, K_2)$. Thus, if $UB(S \cup X, Y) < \delta$, we discard the current search space.

Based on the finding of [31] for frequent itemsets, the complexity of mining exceptional subgraphs is NP-hard. Therefore, we have no guarantee on the execution time of Algorithm 1, as the number of exceptional subgraphs can be exponential in the size of the dataset. However, each recursive call has a worst case time complexity in $O(\max\{|C| \times |V|, |V| + |E|\})$:

- Computing UB on Line 2 or $WRAcc$ on Line 12 take $O(|S \cup X| \times |Y|)$ i.e. $O(|C| \times |V|)$ in the worst case
- The computation of the next candidate $x \in X$ with the fail first principle (Line 4) requires in the worst case $O(|C| \times |V|)$
- Line 5 takes $O(|Y|)$, that is to say $O(|V|)$ at most
- Line 6, computing the connected components, takes $O(|V| + |E|)$
- Line 7, $f(CC)$ is obtained in $O(|C| \times |CC|)$. For all the connected components of $G[Y']$, this requires in overall $O(|C| \times |Y'|)$, which corresponds to $O(|C| \times |V|)$ in the worst case.

CEnergetics enumerates maximal patterns in a depth-first manner. The search space can be represented as a tree where each enumerated maximal pattern (S, Y) corresponds to a single leaf. The depth of this tree is bounded by $2 \times |C|$, since in each recursive call at least one element $x \in X$ is added to S. Thus, the number of recursive calls between two leaves is bounded by $4 \times |C|$ (we backtrack at most $2 \times |C|$ times and then we go in depth at most $2 \times |C|$ times). Thus, we can conclude that the time delay between the enumeration of two leaves of this tree (two different maximal patterns) is polynomial in $O(|C| \times \max\{|C| \times |V|, |V| + |E|\})$.

3.2 The exceptional subgraph space sampling approach

In practice, end-users want to obtain high-quality patterns in a short amount of time, especially in interactive data mining processes. However, we show in

experiments that the runtime of CEnergetics increases when the graph size or the number of attributes increase, and it may require a considerable time to mine very large graphs. To overcome this issue, we propose an approach that computes a sampling of the closed exceptional subgraphs within a user-given time-budget.

We adapt the randomized pattern mining technique of [3] to exceptional subgraphs 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 subgraphs. But this strategy failed in computing patterns with high WRAcc values because the graph structure was neglected. Thus, we adopted the reverse approach that consists in randomly generating maximal patterns (S, K).

We perform a random walk on a graph whose vertices are the maximal patterns and the edges connect couple of patterns (S_1, K_1) and (S_2, K_2) such that $K_1 \subseteq K_2$ and there does not exist a maximal pattern (S, K) such that $K_1 \subset K \subset K_2$ (strict inclusion).

To define how is constructed the graph on which the random walk is performed, we need to introduce two new functions

- $comp : 2^V \times 2^V \rightarrow 2^V$: Given two subsets of vertices H and K such that $K \subseteq H$ and $G[K]$ is connected, $comp(K, H)$ returns the connected component of H that contains K.
- $clo : 2^V \rightarrow 2^V$: Given a connected subgraph induced by K, $clo(K)$ returns the part of the closure of K that is connected and contains K:

$$clo(K) = comp(K, g(f(K)))$$

  $clo(K)$ can be computed by extending K recursively with all neighbors v that maintain $f(K \cup \{v\}) = f(K)$.

During the random walk, edges (transitions) are chosen following a probability measure that favors high-quality patterns:

1. The random walk starts by drawing a first vertex using the probability $\mathcal{P}(\{v\}) = \frac{WRAcc(f(\{v\}), clo(\{v\}))}{\sum_{u \in V} WRAcc(f(\{u\}), clo(\{u\}))}$ to form the first explored maximal pattern $(f(\{v\}), clo(\{v\}))$.
2. A new maximal pattern is generated from the pattern (S, K) by considering all maximal patterns that are direct super-sets of K. Such patterns are generated by alternatively adding a neighbor element $v \in N(K) \setminus K$ to K and considering the closure $clo(K \cup \{v\})$. $N(K)$ is the set of neighbors of K: $N(K) = \{v \in V \mid \exists u \in K : (u, v) \in E\}$. (S, K) is also considered among the patterns that can be generated in the next step. The set $Next(K)$ of all possible next subgraphs is then:

$$Next(K) = \{K\} \cup \{clo(K \cup \{v\}) \mid v \in N(K) \setminus K\}$$

Thus, from $Next(K)$, all the direct successors to (S, K) can be enumerated by:

$$\{(S', K') \mid K' \in Next(K) \text{ and } S' = f(K')\}$$

The next random step is drawn based on the probability $\mathcal{P}(K' \mid K)$, that is the probability to reach $K' \in Next(K)$ from K: $\mathcal{P}(K' \mid K) = \frac{WRAcc(f(K'),K')}{\sum_{K_2 \in Next(K)} WRAcc(f(K_2),K_2)}$. This distribution of probabilities rewards transitions toward maximal patterns with large $WRAcc(f(K'), K')$ value.

3. The current random walk stops when $K' = K$ and a new one is started from step (1). Otherwise, the random walk continues by repeating Step (2) on the set of vertices K'. At each step of the random walk, if $WRAcc(f(K), K) \geq \delta$ and $|K| \geq \sigma$, the pattern is added to the output result set.

The algorithm EXCESS[1] (see Algorithm 2) samples patterns until the specified execution time is consumed. Since K is extended at each iteration by at least one vertex v, and K is bounded by V, the extension loop (Line 9) stops after at most $|V|$ iterations.

Algorithm 2: EXCESS(time_Budget, δ, σ)

```
   Input: time_Budget
   Output: R a set of sampled patterns
 1 for v ∈ V do
 2 |   if WRAcc(f({v}), clo({v})) ≥ δ and |clo({v})| ≥ σ then
 3 |   |   R ← R ∪ (f({v}), clo({v}))
 4 while current_time < time_Budget do
 5 |   // Step 1: draw a vertex v
 6 |   draw v ~ WRAcc(f({v}),clo({v})) / Σ_{u∈V} WRAcc(f({u}),clo({u}))
 7 |   // Step 2: expansion of K
 8 |   K' ← clo({v})
 9 |   repeat
10 |   |   K ← K'
11 |   |   // Compute the set Next(K)
12 |   |   Next(K) ← {K}
13 |   |   for v ∈ N(K) \ K do
14 |   |   |   Next(K) ← Next(K) ∪ {clo(K ∪ {v})}
15 |   |   for K' ∈ Next(K) do
16 |   |   |   if WRAcc(f(K'), K') ≥ δ and |K'| ≥ σ then
17 |   |   |   |   R ← R ∪ (f(K'), K')
18 |   |   draw K' ~ WRAcc(f(K'),K') / Σ_{K2∈Next(K)} WRAcc(f(K2),K2)
19 |   until K' = K;
```

In the following, we prove that all maximal patterns with nonzero $WRAcc$ value have a non zero probability to be generated. To this end, we first prove the following necessary property.

[1] EXCESS stands for EXceptionnal ClosEd Subgraph Sampler.

Property 3 For each maximal pattern $P = (S, K)$ with $|K| \geq 1$, there exists a maximal pattern $P^\star = (S^\star, K^\star)$ s.t: $K^\star \subset K$ (a strict inclusion) and $\exists v^\star \in N(K^\star) \setminus K^\star$ with $K = clo(K^\star \cup \{v^\star\})$.

Proof Since we know that for each maximal pattern $P = (S, K)$ with $|K| \geq 1$ there exists a maximal pattern $P' = (S', K')$ s.t $K' \subset K$ (at least the empty pattern $P' = ((C^+, C^-), \emptyset)$), we prove the property by induction: $\forall n \in \mathbb{N}^\star$, for each maximal pattern $P = (S, K)$ such that there exists a maximal pattern $P' = (S', K')$ with $K' \subset K$ and $|K| - |K'| \leq n$, there also exists a maximal pattern $P^\star = (S^\star, K^\star)$ with $K^\star \subset K$ and $\exists v^\star \in N(K^\star) \setminus K^\star$ with $K = clo(K^\star \cup \{v^\star\})$.

- If $n = 1$: $K \setminus K' = \{v\}$, then $clo(K' \cup \{v\}) = clo(K) = K$. Thus $P^\star = P'$
- Let us suppose that the proposition is true for n. Let $P = (S, K)$ be a maximal pattern for which there exists a maximal pattern $P' = (S', K')$ s.t $K' \subset K$ and $|K| - |K'| \leq n+1$. If $|K| - |K'| \leq n$, then the proposition is verified according the the induction hypothesis. Otherwise $|K| - |K'| = n+1$, let $v \in K \cap N(K') \setminus K'$, since $K' \cup \{v\} \subset K$ then $clo(K' \cup \{v\}) \subseteq K$:
  - If $clo(K' \cup \{v\}) = K$, then $P^\star = P'$
  - If $clo(K' \cup \{v\}) \neq K$. We have $clo(K' \cup \{v\}) \subset K$, and $(f(K' \cup \{v\}), clo(K' \cup \{v\}))$ is a maximal pattern, and $|K| - |clo(K' \cup \{v\})| \leq n$. Then, according to the induction hypothesis, the proposition is verified.

Property 4 For each maximal pattern $P = (S, K)$ with $WRAcc(S, K) > 0$, the probability $\widetilde{\mathcal{P}}(P)$ that the random walk reaches the pattern P is not null: $\widetilde{\mathcal{P}}(P) > 0$.

Proof Let us prove by induction on $n \in \mathbb{N}^\star$, that for all maximal pattern $P = (S, K)$ s.t $WRAcc(S, K) > 0$ with $|K| \leq n$: $\widetilde{\mathcal{P}}(P) > 0$.

- For $n = 1$: $K = \{v\}$, and $K = clo(\{v\})$, P can be sampled directly in Step 1:
$$\widetilde{\mathcal{P}}(P) \geq \frac{WRAcc(S, K)}{\sum_{u \in V} WRAcc(f(\{u\}), clo(\{u\}))} > 0$$
- Let us suppose that the proposition is true for n. Let $P = (S, K)$ be a maximal pattern s.t $WRAcc(S, K) > 0$ and $|K| = n + 1$. According to Property 3, there exists a maximal pattern $P^\star = (S^\star, K^\star)$ s.t: $K^\star \subset K$ and $\exists v^\star \in N(K^\star) \setminus K^\star$ with $K = clo(K^\star \cup \{v^\star\})$. If $K^\star = \emptyset$, then $K = clo(\{v^\star\})$, this means that P can be sampled on Step 1:
$$\widetilde{\mathcal{P}}(P) \geq \frac{WRAcc(S, K)}{\sum_{u \in V} WRAcc(f(\{u\}), clo(\{u\}))} > 0$$
If $K^\star \neq \emptyset$, since $WRAcc(S, K) > 0$, then $S \neq (\emptyset, \emptyset)$, and we know that $S \subseteq S^\star$, thus $S^\star \neq (\emptyset, \emptyset)$. This means that $WRAcc(S^\star, K^\star) > 0$. In the other hand, $K^\star \leq n$, then $\widetilde{\mathcal{P}}(P^\star) > 0$. Also, $K \in Next(K^\star)$. So, P can be reached after sampling $P^\star$:
$$\widetilde{\mathcal{P}}(P) \geq \widetilde{\mathcal{P}}(P^\star) \times \frac{WRAcc(f(K), K)}{\sum_{K_2 \in Next(K^\star)} WRAcc(f(K_2), K_2)} > 0$$

Since each maximal pattern $P = (S, K)$ with $WRAcc(S, K) > 0$ can be reached by the random walk, we can conclude that if $WRAcc(S, K) \geq \delta$ and $|K| \geq \sigma$, then the pattern P has a non zero probability to be returned by EXCESS. Furthermore, the used probability distribution rewards high-quality patterns by giving them more chance to be sampled.

4 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[2].

4.1 Datasets and aims

We considered 10 real-world datasets whose characteristics are given in Table 2. Eight of them come from [8] and depict Foursquare venues over 4 US and 4 EU important cities. The venues are described by a hierarchy[3]. 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[4] 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 2: Description of the real-world datasets

[2] https://github.com/AnesBendimerad/ClosedExceptionalSubgraphMining
[3] https://developer.foursquare.com/categorytree
[4] https://www.kaggle.com/c/sf-crime

In this experimental study, we aim to examine the behaviors of CEnergetics and EXCESS regarding the following questions:

- What is the efficiency of CEnergetics with regard to the graph characteristics that may affect its execution time?
- How effective are CEnergetics' pruning properties?
- Does CEnergetics scale?
- Does EXCESS provide a good sample of Exceptional subgraphs?
- What about the relevancy of Exceptional subgraphs?

No related work, among those presented in Section 5, can be used as a competitor of CEnergetics. Indeed, algorithms of pattern extraction in vertex attributed graphs [22,10,23,29,26,4] compute dense subgraphs whose vertices have homogeneous attribute values, while CEnergetics focuses on subgraphs whose vertex attributes are different from those of the rest of the graph. Other related works, that look for exceptional subgraphs [12,18], are designed for graphs with attributes on the edges. Thus, in this section, we compare our two novel algorithms only to the ones of our first attempt [2]: We demonstrate that CEnergetics is more efficient than Energetics (a complete algorithm that extracts non closed exceptional subgraphs) and is able to tackle graphs with more than 150 attributes while Energetics fails with 50 attributes. Furthermore, our new pattern sampler algorithm EXCESS provides better results than EXPreSS. Finally, we report some examples of exceptional subgraphs on real-world data and discuss the insights they convey.

4.2 Quantitative study

We compare the efficiency and the effectiveness CEnergetics and Energetics according to the number of attributes and the number of vertices. To this end, we consider the New York graph described in Table 2. We vary the number of vertices and attributes by removing or duplicating vertices and attributes. Figure 4 reports the runtime, the number of explored patterns and the number of returned patterns of both CEnergetics and Energetics on this testbed. The values of parameters are: $\delta = 0.01$, $\sigma = 1$. CEnergetics clearly has an advantage over Energetics. It is much faster, explores a lower number of candidates, and return a much more concise set of patterns. The differences between the two algorithms are more important when the number of attributes varies. CEnergetics outperforms Energetics with several order of magnitudes. Furthermore, CEnergetics is able to handle graphs with more than 150 attributes while Energetics fails as soon as graphs involve more than 40 attributes.

We now focus on the study of CEnergetics with respect to the parameters of the algorithm (i.e., σ, the minimum number of vertices involved in a pattern, and δ the minimum WRAcc threshold). By default, these values are set to $\delta = 0.01$ and $\sigma = 1$ in order to not being stringent. Fig. 5 reports the behavior (i.e., runtime, number of explored sub-graphs and number of

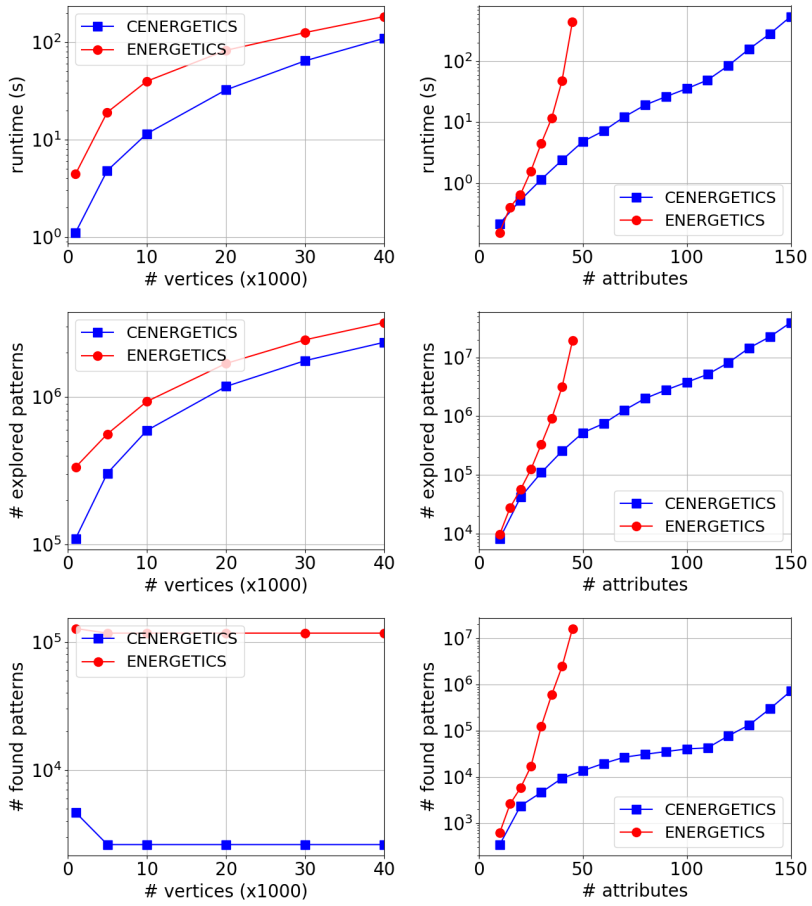


Fig. 4: Comparison (i.e., runtime, number of explored patterns and number of returned patterns) between CENERGETICS and ENERGETICS according to the number of vertices and attributes (default values: number of vertices = 1000, number of attributes = 30).

patterns) of CENERGETICS on the 10 real-world datasets when varying the input parameters δ and σ. The obtained results confirm the previous findings. The execution time and the numbers of explored and returned patterns increase when the thresholds become less stringent. Interestingly, *S.F. C&V* is the dataset whose execution times are the most important. This confirms the previous finding that the number of attributes is the most influential data parameter in the discovery of exceptional subgraphs.

We also study the behaviour of our algorithm with regard to the replication factor. For a replication factor equal to n, the attributed graphs are duplicated n times such that the initial vertices are repeated n times with the same attributes values and the same connections with the corresponding duplicated

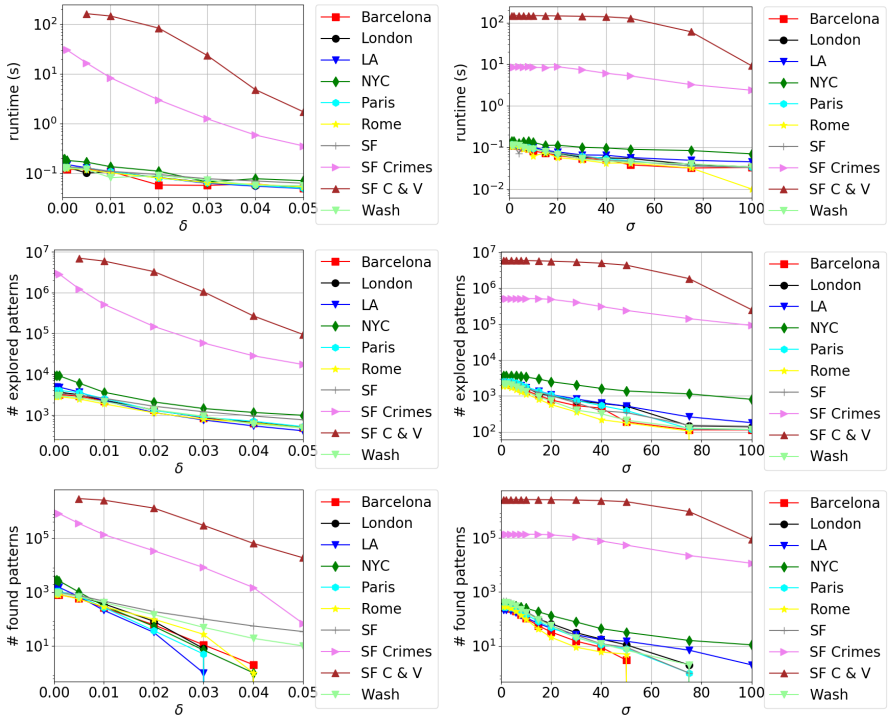


Fig. 5: Behavior of CENERGETICS (runtime in 1st raw, #explored patterns in 2nd raw, #patterns in 3rd raw) according to δ (1st column) and σ (2nd column) for the 10 real-world datasets (default values: $\delta = 0.01, \sigma = 1$).

vertices. Therefore, a n-duplicated attributed graphs correspond to n identical attributed graphs that are not connected together and thus contains n times the number of exceptional subgraphs of the original graph. For each replicate attributed graph, we compute the ratio of the execution time of CENERGETICS on the duplicated graph to the execution time of CENERGETICS on the original graph. Figure 6 reports this ratio for the 10 replicated graphs. For most of the datasets, the algorithm behaves almost linearly with respect to the replication factor. However, this is not the case for *S.F. C&V* and *S.F. Crimes* that are the datasets with the highest number of attributes. For these two datasets, the performance degrades when the replication factor increases. The runtime ratio increases superlinearly with the replication factor.

In order to demonstrate the effectiveness of the pruning techniques used (the upper bound UB, and the Fail First Principle FFP), we compare the performance of CENERGETICS in four different configurations:

1. no opt: in this configuration, none of the pruning techniques is used.
2. FFP: we only use the Fail First Principle (FFP).
3. UB: we only use the upper bound UB.

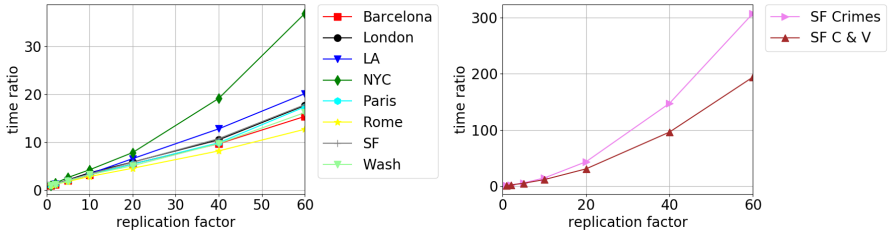


Fig. 6: Runtime ratio with respect to the replication factor for real world datasets ($\delta = 0.01$ except SF Crimes and SF C&V (0.03 and 0.05), $\sigma = 1$).

4. UB+FFP: we use both UB and FFP.

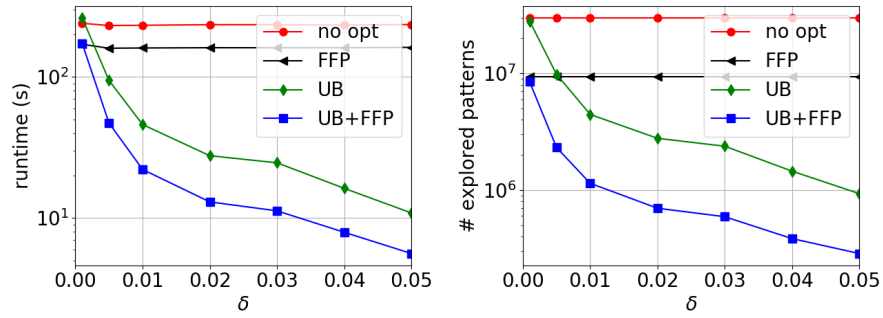


Fig. 7: Impact of pruning techniques on runtime (1st column), and the number of explored patterns (2nd column). The number of discovered patterns decreases from 10^6 to 140 (not reported on the figures).

We performed these four configurations on an attributed graph involving 10000 vertices and 30 attributes built by duplicating the NYC Foursquare graph. It is important to note that *no opt* configuration considers closed exceptional subgraphs which makes the extraction feasible. We study the runtime and the number of explored sub-graphs when varying the value of δ. Results are given in Figure 7. UB+FFP outperforms all the other configurations with at least one order of magnitude, especially when the value of δ is increased. Indeed, the use of UB takes benefit from the minimum threshold δ in order to reduce the runtime and the number of explored patterns. These results confirm that even if UB is the most effective technique, the simultaneous consideration of UB and FFP makes the algorithm much more efficient.
These first experiments demonstrate that CENERGETICS is only efficient for graphs whose number of attributes is rather small (at most 150). Indeed, CENERGETICS is not able to manage attributed graphs with large number of attributes (e.g., hundreds). EXCESS has been designed especially to per-

form 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 EXCESS to compute exceptional subgraphs of high $WRAcc$ values, we report in Figure 8 the distributions of the WRAcc measure of both the complete set of exceptional subgraphs returned by CEnergetics and the sample provided by EXCESS. Several time budgets are used and they are all lower than the execution time required by CEnergetics. We can observe that the two distributions are similar and the sampling approach succeeds in fostering patterns with high WRAcc measure. Also, the higher the time budget, the better the distribution. Figure 9 reports similar distributions for the real-world datasets with hundreds of attributes for which an exhaustive search is not possible. The distributions are similar. Thus, EXCESS makes it possible to discover high quality patterns within a time-budget.
We also compare EXCESS with EXPreSS [2] which does not take into account closed patterns. Distributions of patterns sampled by each of these approaches are reported in Figure 10 using a logarithmic scale. These results reveal that EXCESS returns a larger sampling than EXPreSS for the same time budget. Interestingly, EXCESS provides much more patterns with higher $WRAcc$ values than patterns sampled by EXPreSS. This confirms that EXCESS is able to extract more patterns of better quality (i.e., with high $WRAcc$ values) than EXPreSS.

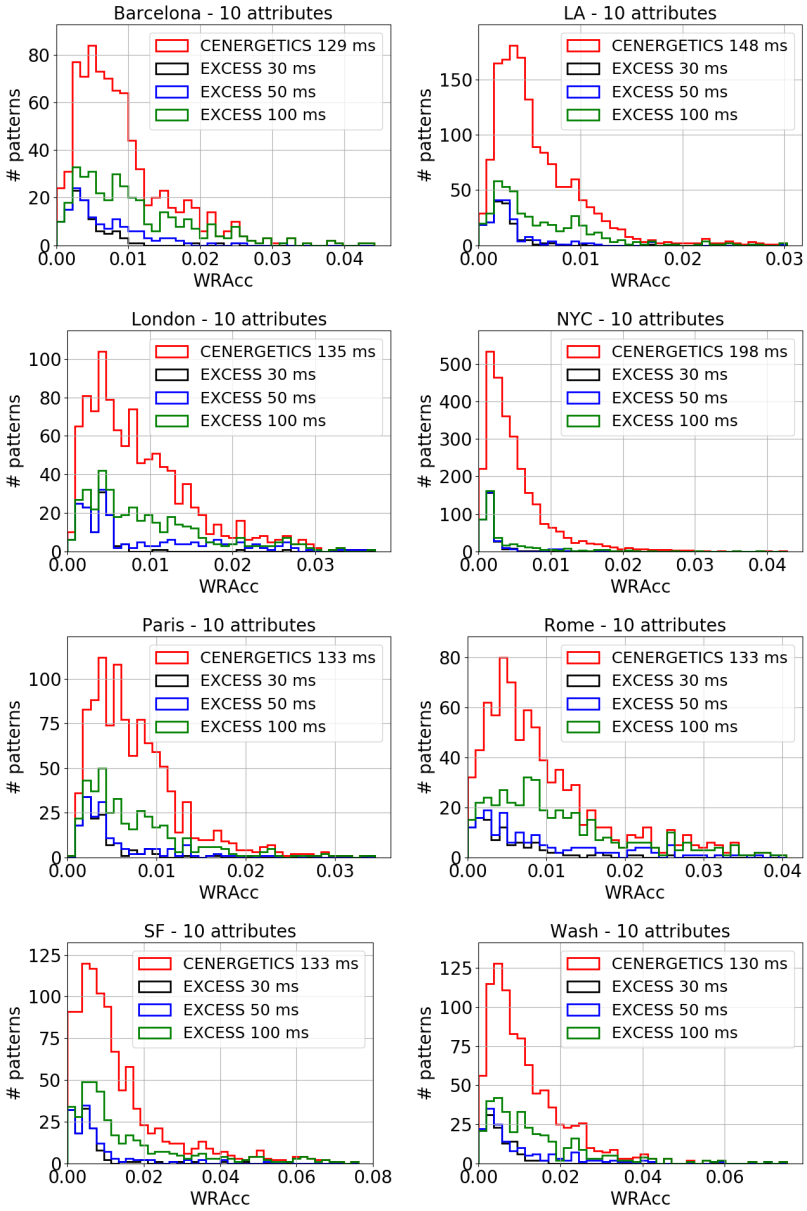


Fig. 8: Distributions of the patterns from CENERGETICS and EXCESS with different time budgets ($\delta = 0$). The number of attributes is 10.

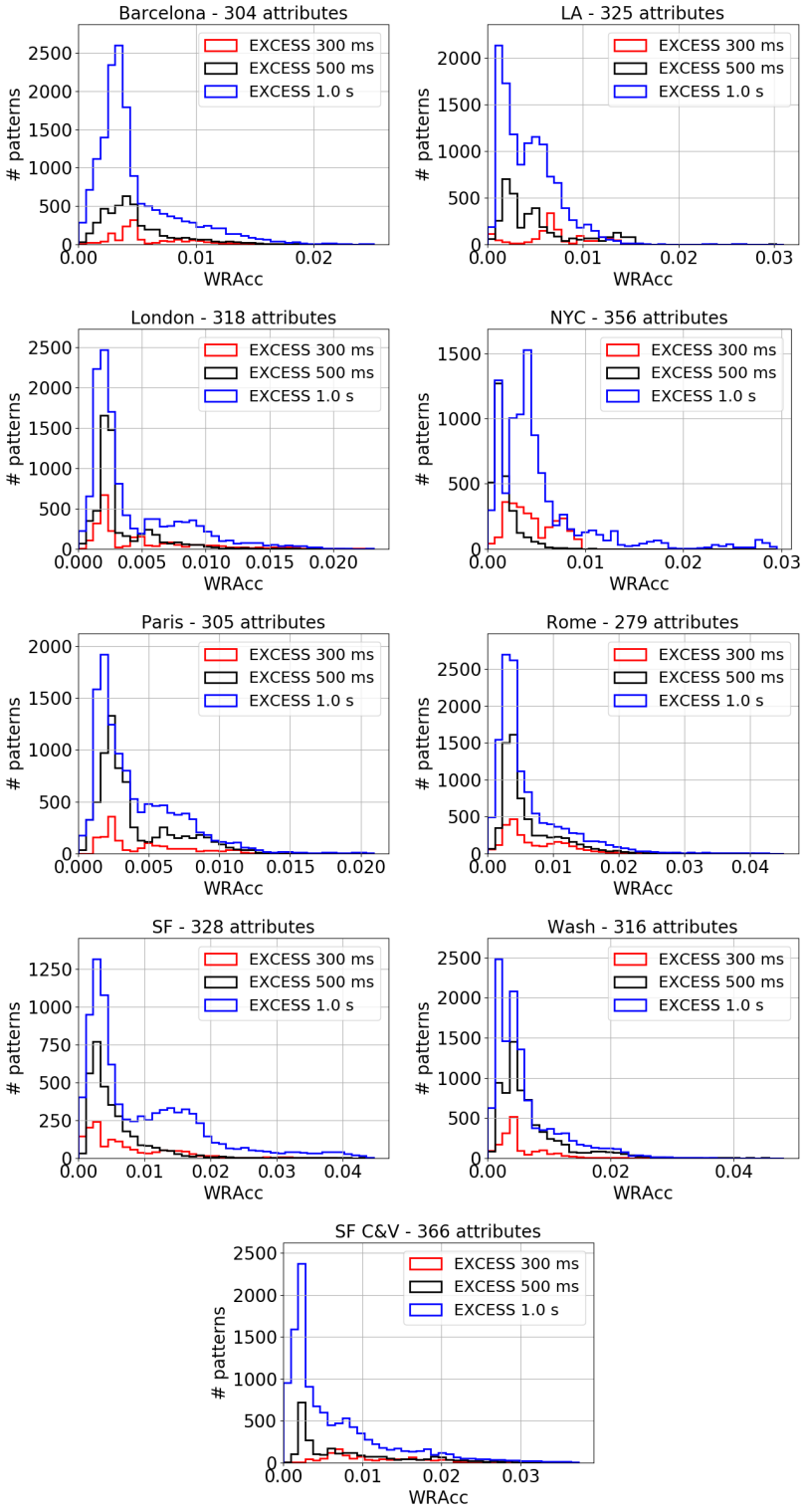


Fig. 9: Distribution of the output space sampling with different time budgets for datasets with larger number of attributes.

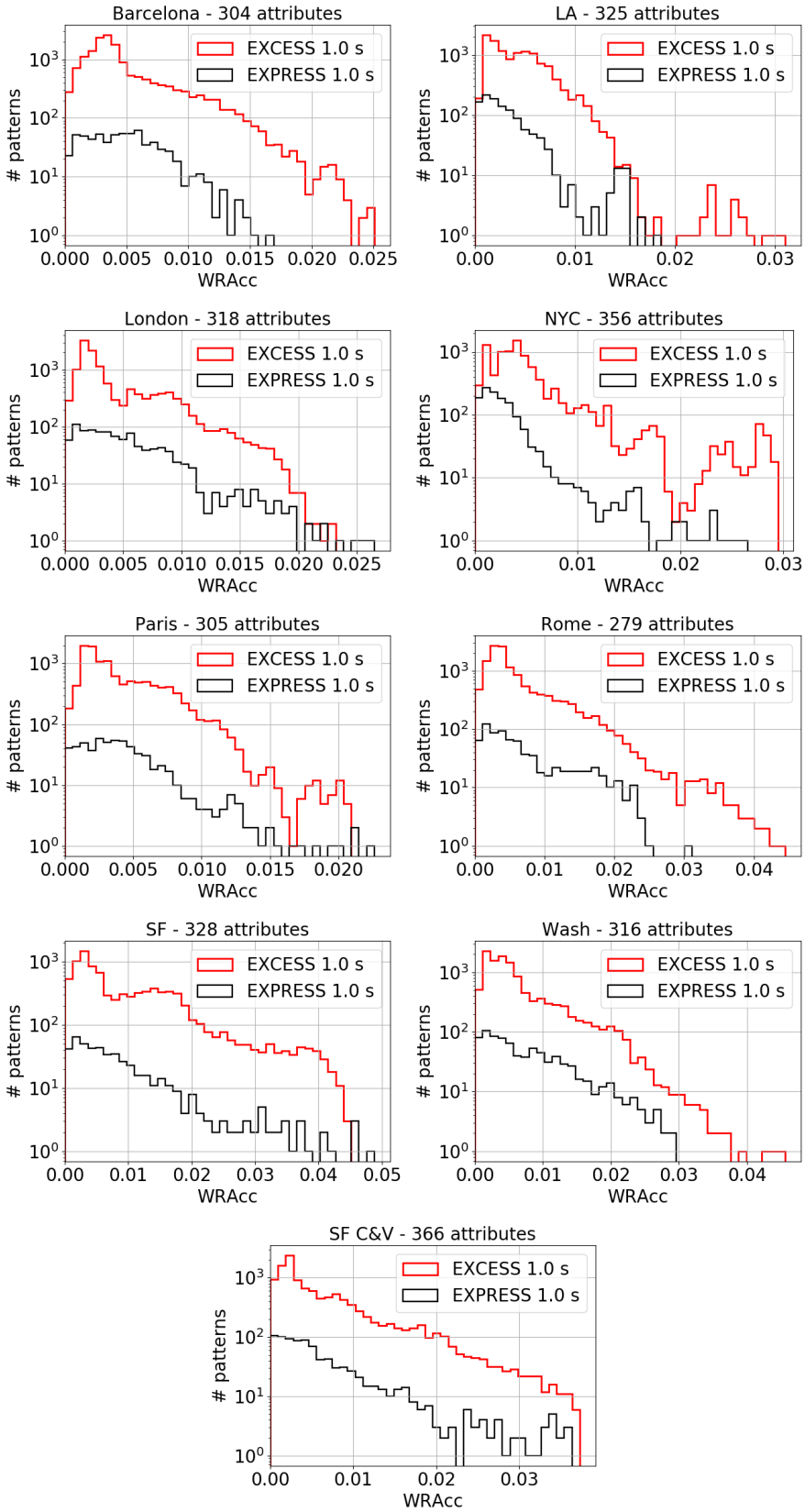
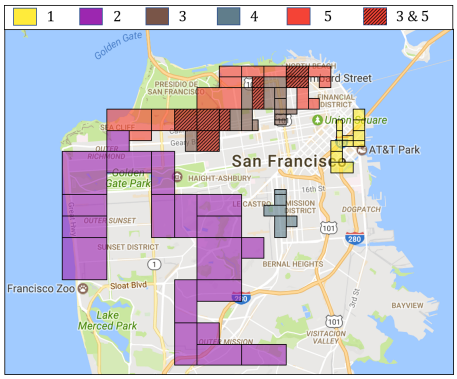


Fig. 10: Comparison of distributions of patterns sampled with EXCESS and EXPRESS

4.3 Qualitative study



Id	S^+ Venues	S^+ Crimes	$WRAcc$	Size
P_1	Professional & other places	Embezzlement	0.007	11
P_2	Residence	Vandalism, Vehicle theft, Burglary	0.006	19
P_3	Shop & services	Burglary	0.06	21
P_4	Shop & services, Nightlife spot, Art & Entertainment, Food	Drunkenness	0.003	5
P_5	Outdoors & recreation	Burglary	0.002	20

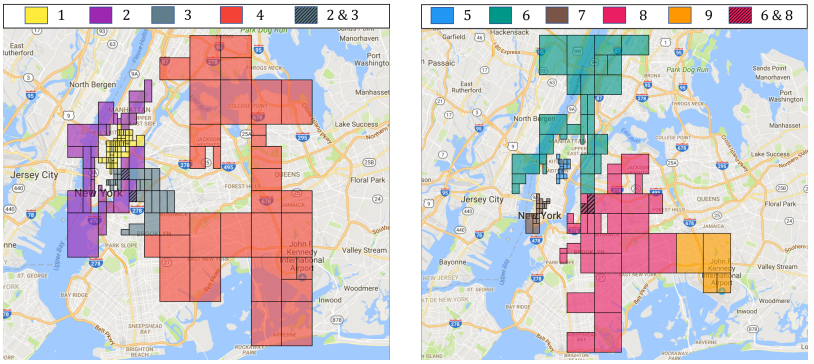
Fig. 11: Patterns discovered in San Francisco crimes and venues dataset (49 attributes)

We use CENERGETICS on San Francisco crimes and venues dataset to automatically identify typical areas of this city. Figure 11 displays 5 discovered patterns. Pattern P_1 depicts neighbourhoods with a high concentration of venues of type professional & other places, and crimes of type embezzlement. This can be explained by its proximity to the Financial District. P_2 is a neighborhood located in the West and South-West of San Francisco. It contains a positive contrast of residences and crimes of type vandalism, vehicle theft, burglary. These crimes are known as the most common types of crimes in residential areas. P_3 and P_5 are overlapping patterns located in the North of the city. They characterize areas with a high concentration of venues of type shop & services, outdoors & recreation, and crimes of type burglary. Pattern P_4 describes a neighborhood with a positive contrast of crimes related to drunkenness, which can be explained by the high concentration of nightlife spots in this area.

We also report 9 discovered patterns on New York venues dataset. They are presented in Fig. 12. Four of them (on the left-hand side map) are discovered on the dataset with 10 attributes, whereas the 5 remaining ones (on the right-hand side map) are discovered on the dataset with 356 attributes. P_1 is located in the South of Central Park. This neighborhood is known to be a business and professional area with a low concentration of residences. A sub-area of P_1 is depicted by P_5 with a high concentration of offices, buildings, medical centers. P_2 describes areas with a high proportion of venues of

type outdoors & recreation. It contains Central Park and some areas located near East River and Hudson River. P_3 covers a part of the South of Manhattan and the North of Brooklyn, with a high concentration of nightlife spots. P_4 covers John F. Kennedy and LaGuardia Airports and their surroundings. This explains the high presence of travel & transport venues. More precisely, P_9 contains neighborhoods of John F. Kennedy Airport, and it depicts them with venues of types: Taxi, parkings, donut shops, airport, and general travels. Both P_6 and P_8 represents areas with high proportion of residences. P_8 is also characterized with an important concentration of food & drink shops. P_7 is another pattern that describes a part of South Manhattan with a high concentration of offices.

Besides, we mined exceptional subgraphs 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. The higher concentration of outdoor & recreation places is surrounding for London. For seaside towns, they are concentrated on the coasts.



Id	S^+	S^-	$WRAcc$	size
P_1	Professional & other places	Residence	0.039	50
P_2	Outdoors & recreation	Shop & services, Food	0.023	35
P_3	Nightlife spot	College & university, Travel & transport, Professional & other places	0.02	27
P_4	Travel & transport	Art & entertainment	0.016	32
P_5	Office, Building, Medical center		0.01	16
P_6	Residential building (apartment / condo), Home (private)		0.008	30
P_7	Office		0.008	17
P_8	Home (private), Food & drink shop		0.007	30
P_9	Taxi, Parking, Donut shop, Airport, General travel		0.006	4

Fig. 12: Patterns discovered in New York datasets with 10 attributes (patterns P_1 to P_4 plotted on the left-hand side map) 356 attributes (patterns P_5 to P_9 plotted on the right-hand side map).

5 Related work

Several approaches have been designed to discover new insights in vertex attributed graphs. The pioneering work of Moser et al. [22] presents a method to mine dense homogeneous subgraphs, i.e., subgraphs whose vertices share a large set of attributes. Similar to that work, Günnemann et al. [10] introduce a method based on subspace clustering and dense subgraph mining to extract non redundant subgraphs that are homogeneous with respect to the vertex attributes. In [23], the authors aim to discover collections of maximal cliques whose vertices are homogeneous with respect to the vertex attributes. Silva et al. [29] extract pairs made of a dense subgraph and a Boolean attribute set such that the Boolean attributes are strongly associated with the dense subgraphs. In [26,4], the authors propose to mine the graph topology of a large attributed graph by finding regularities among numerical vertex descriptors. In [27], the authors compute subgraphs that maximize an objective function on a vertex and edge weighted graph, seen as an activity graph by assuming a single numerical attribute on the graph nodes. The main objective of all these approaches is to find regularities instead of peculiarities within a large graph, whereas *Exceptional subgraph Mining* computes subgraphs 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. In [12], the authors aim at discovering contextualized subgraphs that are exceptional with respect to a model of the dataset. Restrictions on the attributes, that are associated to edges, are used to generate subgraphs. Such patterns are of interest if they pass a statistical test and have high value on an adapted WRAcc measure. Similarly, [18] propose to discover subgroups with exceptional transition behavior which is assessed by first-order Markov chain model. More generally, Subgroup Discovery [14,24] 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* [17,16,5]. A variety of measures of changes have been used: pairwise correlation and entropy measures [17], Bayesian networks [6], Kullback-Leibler Divergence and encoding difference based on Minimum Description Length [15]. 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 for a special purpose sampling procedure

tailored for a specific set of itemset mining tasks [3,21,20,9,19]. Interestingly, [7] propose to take benefit from the latest advances in sampling solutions in SAT to define a flexible (w.r.t the choice of constraints and sampling distributions) pattern sampling algorithm with theoretical guarantees regarding sampling accuracy. However, this framework only supports pattern languages that can be compactly represented with binary variables, such as itemsets. Numeric attributes cannot be handled without discretization. Also, this approach has not been applied to richer pattern languages yet. Some researchers have tackled the problem of sampling the output space of frequent subgraphs in a collection of graphs [11,28]. These methods are based on random walks. In particular, [28] 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. Besides, these methods aim at sampling frequent subgraphs while we address the problem of exceptional attributed subgraph sampling which is much more challenging since we have to deal simultaneously with two dimensions: Subgraphs and characteristics. Our approach is based on a random walk over the closed subgraphs that fosters patterns with a high WRAcc measure.

6 Conclusion

We introduced the closed exceptional subgraph mining problem to discover homogeneous areas that differ from the rest of the city. We defined an efficient algorithm that computes the complete set of exceptional subgraphs by taking advantage of a closure operator, a tight upper bound and other pruning properties. Focusing on closed pattern avoids redundancy among exceptional subgraphs. We also designed an algorithm to sample the output space of closed patterns to enable time-budget analysis. We reported an extensive empirical study over 10 real-world datasets that demonstrates the relevancy of our proposal. Experiments give evidence about the efficiency and the effectiveness of CEnergetics that outperforms it previous version Energetics with several order of magnitudes. However, CEnergetics still has difficulties to scale with the number of attributes. This problem is fixed by EXCESS 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.

We believe that this work opens new directions for future work. For instance, the simultaneous consideration of a collection of cities makes it possible to highlight the similarities and the differences between them. Our proposal can be extended to take into account other graph topological properties (e.g., diameter, reachability) and other quality measures can be investigated to highlight some phenomena over the city areas like outliers or anomalies. Another interesting direction is the interactive discovery of exceptional subgraphs in urban data. To this end, an instant mining approach (i.e., pattern sampler)

has to be coupled to the learning of a user model based on her feedback to foster the interactive process.

Acknowledgements 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.

References

1. Atzmüeller, M., Doerfel, S., Mitzlaff, F.: Description-oriented community detection using exhaustive subgroup discovery. Information Science **329**, 965–984 (2016)
2. Bendimerad, A.A., Plantevit, M., Robardet, C.: Unsupervised exceptional attributed sub-graph mining in urban data. In: IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain, pp. 21–30 (2016)
3. Boley, M., Lucchese, C., Paurat, D., Gärtner, T.: Direct local pattern sampling by efficient two-step random procedures. In: ACM SIGKDD 2011, pp. 582–590 (2011)
4. Boulicaut, J., Plantevit, M., Robardet, C.: Local pattern detection in attributed graphs. In: Solving Large Scale Learning Tasks. Challenges and Algorithms - Essays Dedicated to Katharina Morik on the Occasion of Her 60th Birthday, pp. 168–183 (2016)
5. Duivesteijn, W., Feelders, A., Knobbe, A.J.: Exceptional model mining - supervised descriptive local pattern mining with complex target concepts. Data Mining and Knowledge Discovery **30**(1), 47–98 (2016)
6. Duivesteijn, W., Knobbe, A.J., Feelders, A., van Leeuwen, M.: Subgroup discovery meets bayesian networks – an exceptional model mining approach. In: ICDM 2010, pp. 158–167 (2010)
7. Dzyuba, V., van Leeuwen, M., Raedt, L.D.: Flexible constrained sampling with guarantees for pattern mining. Data Mining and Knowledge Discovery **(accepted)** (2017)
8. Falher, G.L., Gionis, A., Mathioudakis, M.: Where is the soho of rome? measures and algorithms for finding similar neighborhoods in cities. In: ICWSM 2015, pp. 228–237 (2015)
9. Giacometti, A., Soulet, A.: Frequent pattern outlier detection without exhaustive mining. In: PAKDD 2016, pp. 196–207 (2016)
10. Günnemann, S., Färber, I., Boden, B., Seidl, T.: Subspace clustering meets dense subgraph mining. In: ICDM 2010, pp. 845–850 (2010)
11. Hasan, M.A., Zaki, M.J.: Output space sampling for graph patterns. PVLDB **2**(1), 730–741 (2009)
12. Kaytoue, M., Plantevit, M., Zimmermann, A., Bendimerad, A., Robardet, C.: Exceptional contextual subgraph mining. Machine Learning pp. 1–41 (2017)
13. Kuznetsov, S.O.: Learning of simple conceptual graphs from positive and negative examples. In: Principles of Data Mining and Knowledge Discovery, Third European Conference, PKDD '99, Prague, Czech Republic, September 15-18, 1999, Proceedings, pp. 384–391 (1999)
14. Lavrac, N., Kavsek, B., Flach, P.A., Todorovski, L.: Subgroup discovery with CN2-SD. Journal of Machine Learning Research **5**, 153–188 (2004)
15. van Leeuwen, M.: Maximal exceptions with minimal descriptions. Data Mining Knowledge Discovery **21**(2), 259–276 (2010)
16. van Leeuwen, M., Knobbe, A.J.: Diverse subgroup set discovery. Data Mining Knowledge Discovery **25**(2), 208–242 (2012)
17. Leman, D., Feelders, A., Knobbe, A.J.: Exceptional model mining. In: ECMLPKDD 2008, pp. 1–16 (2008)
18. Lemmerich, F., Becker, M., Singer, P., Helic, D., Hotho, A., Strohmaier, M.: Mining subgroups with exceptional transition behavior. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, pp. 965–974 (2016)

19. Li, G., Zaki, M.J.: Sampling frequent and minimal boolean patterns. Data Mining Knowledge Discovery **30**(1), 181–225 (2016)
20. Moens, S., Boley, M.: Instant exceptional model mining using weighted controlled pattern sampling. In: IDA, pp. 203–214 (2014)
21. Moens, S., Goethals, B.: Randomly sampling maximal itemsets. In: SIGKDD Workshop on Interactive Data Exploration and Analytics, pp. 79–86. ACM (2013)
22. Moser, F., Colak, R., Rafiey, A., Ester, M.: Mining cohesive patterns from graphs with feature vectors. In: SDM 2009, pp. 593–604 (2009)
23. Mougel, P., Rigotti, C., Plantevit, M., Gandrillon, O.: Finding maximal homogeneous clique sets. Knowledge Information System **39**(3), 579–608 (2014)
24. Novak, P.K., Lavrac, N., Webb, G.I.: Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. Journal of Machine Learning Research **10**, 377–403 (2009)
25. Park, S., Bourqui, M., Frías-Martínez, E.: Mobinsight: Understanding urban mobility with crowd-powered neighborhood characterizations. In: IEEE International Conference on Data Mining Workshops, ICDM (demo) 2016, December 12-15, 2016, Barcelona, Spain., pp. 1312–1315 (2016)
26. Prado, A., Plantevit, M., Robardet, C., Boulicaut, J.: Mining graph topological patterns: Finding covariations among vertex descriptors. IEEE TKDE. **25**(9), 2090–2104 (2013)
27. Rozenshtein, P., Anagnostopoulos, A., Gionis, A., Tatti, N.: Event detection in activity networks. In: KDD, pp. 1176–1185 (2014)
28. Saha, T.K., Hasan, M.A.: A sampling based method for top-k frequent subgraph mining. Stat. An. & DM **8**(4), 245–261 (2015)
29. Silva, A., Jr., W.M., Zaki, M.J.: Mining attribute-structure correlated patterns in large attributed graphs. PVLDB **5**(5), 466–477 (2012)
30. Spielman, S.E., Thill, J.: Social area analysis, data mining, and GIS. Comp. Env. & Urb. Sys. **32**(2), 110–122 (2008)
31. Yang, G.: The complexity of mining maximal frequent itemsets and maximal frequent patterns. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004, pp. 344–353 (2004)