Contributions to Pattern Mining in Augmented Graphs

Marc Plantevit
Université Claude Bernard Lyon 1
LIRIS UMR5205
HDR Defence, December 14th
Who am I?

PhD in Computer Science, Université Montpellier 2, « Multidimensional Sequence Mining » (2008)


Since 2009: Associate Professor, Université Claude Bernard Lyon 1, LIRIS UMR5205, Data Mining and Machine Learning research group

PC Membership: ≈ 50 conferences (ECMLPKDD, SIAM DM, IDA, IJCAI, AAAI, etc.)
Reviewer for journals: ≈ 10 reviews per years (Dami, Mach, TKDE, IS, SADM, etc.)

Databases, Data Mining, Game Theory, Student R&D projects (TER)
BCS and Master (Data Science, Web and Technologies, AI, Bioinformatics, Theoretical computer science) Students
KDD is not a hype, but a natural evolution of science


Each discipline has grown a theoretical component.

- Babylonian mathematics
- Ancient Egypt: no theorization of algorithms.

Inability to find closed form solutions

Data Mining as a major challenge!

The Fourth paradigm: Data-Intensive Scientific Discovery

Flood of data: new scientific instruments and simulations.

Ability to economically store and manage huge data

The Fourth paradigm:

The Fourth paradigm:

Empirical Science  Theoretical Science  Computational Science  Data Science

1600  1950s  1990s

M. Plantevit, Pattern Mining in Augmented Graphs
Machine Learning or Data Mining? What are the differences?

• Predictive **global** modeling
  • Turn the data into an as **accurate** as possible prediction machine
  • Ultimate purpose is **automatization**
  • E.g., self driving car, pattern recognition

• Exploratory data analysis
  • Automatically discover new **insights** about the domain in which the data was measured
  • Use machine discoveries to synergistically **boost human expertise**
  • E.g., understanding the factors of the olfactory process

Accuracy of Models

Interpretability of results

M. Boley, www.realkd.org

M. Plantevit, Pattern Mining in Augmented Graphs
A good prediction machine does not necessarily provide explicit insights into the data domains

Global linear regression model

Gaussian process model

\[ y = \sum_{j \in S} \alpha_j k(x, x(j)) \]
Global Modelling: the need to explain all the data

- «A complex theory of everything might be of less value than a simple observation about a specific part of the data space»

- Identifying interesting subspace and the power of saying «I don’t know for other points»
Global Modelling is guided by the global picture and may not uncover some interesting insights while local patterns make it possible.

Decision tree with 0.7 accuracy

\[ q_1 = (x_1(i) \in [0.24, 0.71] \land x_2 \geq 0.9) \]

Two subgroups with 1.0 accuracy

\[ q_2 = (x_1(i) \in [0.26, 0.7 \land x_2(i) \leq 0.1) \]
From an Inductive DB Perspective

\( Th(L, D, C) = \{ \psi \in L \mid C(\psi, D) \text{ is true} \} \)

Imielinski and Mannila,
Comm. of the ACM, 1996

- \( L \) a pattern language
- \( D \) a database
- \( C \) some constraints

Pattern mining algorithm:
**enumerating** the elements of
the language that fulfil the
**constraints** within the **data**.

Some variants:
- **Elements**: complete set, top k, representative sample, etc.
- **Constraints**: from satisfaction problem to optimization problem.
- **Data**: {batch data, streaming data} x {graph(s), tree(s), sequence(s), [numerical] itemset(s), etc.}.

Study of the constraint properties to devise effective pruning strategies.
Fill the gap between the user and her data

**Complexity of the data:**
Numerical attributes, graphs, sequences, images, streams vs batches, etc.

**Complexity of the user:**
Explicit interest via constraints or preferences, implicit interest to be learned

**Complexity of the domain:**
How to handle domain knowledge and do not return already known phenomena

**Complexity of the output:**
Do not overwhelm the user! each pattern has an assimilation cost.

Applications
Biology, Neurosciences, Material Engineering, Social Science, Geology, Chemistry, Industry 4.0, …
Some contributions

**Sequence mining**

- Multidimensional sequences (*Plantevit et al., ACM TKDD 2010*)
- Extending some concepts from sequences
  - Strong frequency (*Plantevit and Crémilleux, IDA’09*)
  - δ-free sequences
  - For text mining (*Cellier et al., J. Biomedical Semantics 2015*)

Making the constraint-based pattern mining framework easier to use

- Explicit preferences to overcome the thresholding problem with the skypatterns (*Soulet et al., ICDM’11*), (*Ugarte et al., Artificial Intelligence 2017*)

New pattern domain to analyse rating or vote data

(*Belfodil et al., ECMLPKDD’17*)

Results are **difficult** to **assimilate** for the end-users

Looking for pattern language **easier** to understand
Graphs as a powerful mathematical tool to model real-world phenomena

A vertex: an entity
- user in SN
- scientists
- bicycle stations
- city area
- airport
- (pi|vo)xel
- ...

An edge: an interaction
- friendship/follow
- co-authorship/citations
- travels
- Connectivity

Temporal information (dynamic graphs):
Appearance of vertices/edges through time

Attributes on vertices to better describe the entities
- age, gender, etc.
- distribution of place types
- inner activity

Attributes on edges to better describe the interactions
- weather conditions
- day of week

Graphs are often augmented with additional information

Focus on relational graphs
Overview of the contributions to pattern mining in augmented graphs
Association rules in dynamic graphs
• Revisiting the confidence measure

Trends in dynamic attributed graphs
Definition of the co-evolution patterns and some primitives (size, volume, diameter, similarity, etc.)

Step back:
• A semantics too difficult to apprehend
• Needs deeper investigation to be fully usable in practice.

Vertices (fulfilling some properties) that follow the same trends on some attributes at the same timestamps.

Interestingness based on outside densities.
Taking into account hierarchies on the vertex attributes.

Applications:
Soil erosion, transportation network analysis.
ANR FOSTER

Kim-Ngan Nguyen et al., SDM’11, IDA J 2013

Desmier et al., DS’12, ECMLPKDD’13, IDA’14
Link between structure and vertex attributes

How the graph structure impact the vertex attributes?
Are the vertex attributes correlated with the vertex role within the graph?

Topological patterns
• Based on a generalization of the Kendall’s Tau (Calders et al., KDD’06)
• Rank-correlation between vertex attributes and topological attributes
• The higher the number of publications in Dami, KAIS, and EGC, the lower the Morik number.

Prado et al., IEEE TKDE 2013

Triggering patterns
• Discovering the sequences of vertex attribute variations that impact the structure of the graphs
• <ICDE+, Journal++> → betweenness ++

Kaytoue et al., SNAM 2015

Two Pattern domains to provide insights to these questions.
Step back: Do we fill the gap between the user and her data?

- Pattern domains to discover **new insights**
- Complexity of the data

- Big effort (post-processing of the pattern collection) for the user to find the good knowledge nuggets

- Complexity of the user
- Complexity of the domain
- Complexity of the output
EMM/SD as a basis to handle all the complexities (data, user, domain, output)

Search strategies: beam search, exhaustive search, ...

Subgroup Discovery

Object description

Target concept

target variable(s)

model

Model

Targets

X Y

Models:
Classifier, Bayesian network encoding, ...

Quality measures:
pairwise correlations, entropy, MDL-based measures, WKL, Cook distance, ...

Quality measure

W Klösgen 1996
Leman et al., ECMLPKDD’08
EMM/SD for graphs: challenges

• Which targets and models?
• What about the complexities?
  • Domain
  • User
  • Output

Graphs and/or attributes
Into the model
Biased measures according to feedback
Subjective interestingness

Edge-attributed graphs
Vertex-attributed graphs
Exceptional subgraphs in edge attributed graphs

Kaytoue et al., Machine Learning 2017
Bendimerad et al., Complex Network’18
Aims

**Edge-attributed subgraphs**

\[ G = (V, E, T, \text{Edge}) \]
\[ \text{Edge}: T \rightarrow E \]

- A set of transactions is associated to an edge.
- Alternatively:
  - Several edges between pair of vertices.
  - Each edge is provided with a context that depicts the interaction.

\[ \{m_{14}, m_{15}, m_{16}, m_{17}, m_{18}\} \]
\[ \{m_{12}, m_{13}\} \]
\[ \{m_{9}, m_{10}, m_{11}\} \]
\[ \{m_6, m_7, m_8\} \]

Objectives

Discover a subgraph and a context such that the subgraph is exceptional according to the context.

- Which model to capture it?
- Which quality measure?
- Which pattern language?
- How to extract the patterns?

Exceptional contextual subgraphs

<table>
<thead>
<tr>
<th>Time</th>
<th>Weather</th>
<th>Gender</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day</td>
<td>Rainy</td>
<td>F</td>
<td>20</td>
</tr>
</tbody>
</table>

\( m_1: \)

Age \( \in [20,23] \),
\( \text{Time} = \text{Night} \)
Model: contextual graphs

Context *

<table>
<thead>
<tr>
<th>Time</th>
<th>Weather</th>
<th>Gender</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
E & \rightarrow \{m_{12}, m_{13}\} & C & \rightarrow \{m_6, m_7, m_8\} \\
D & \leftarrow \{m_{14}, m_{15}, m_{16}, m_{17}, m_{18}\} & B & \rightarrow \{m_9, m_{10}, m_{11}\} \\
A & \rightarrow \{m_1, m_2, m_3, m_4, m_5\} & \end{align*}
\]

\[W_s(a, b) = |\{m_1, m_2, m_3, m_4, m_5\}|\]

Context C

<table>
<thead>
<tr>
<th>Time</th>
<th>Weat.</th>
<th>Gend.</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Night</td>
<td>*</td>
<td>*</td>
<td>[20-23]</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
E & \rightarrow \{m_{12}, m_{13}\} & C & \rightarrow \{m_6, m_7, m_8\} \\
D & \leftarrow \{m_{14}, m_{15}, m_{16}, m_{17}, m_{18}\} & B & \rightarrow \{m_9, m_{10}, m_{11}\} \\
A & \rightarrow \{m_2\} & \end{align*}
\]

\[W_C(a, b) = |\{m_2\}|\]

Which edges are surprising?
Quality measure to measure the exceptionality

\( \text{WRAcc} \) measure to assess the exceptionality of \( C \) for an edge \( e \):

\[
\text{WRAcc}(C, e) = \frac{1}{|T|} \left( W_C(e) - \overline{W}_C(e) \right)
\]

\( \overline{W}_C(e) \) is the expected weight: \( \overline{W}_C(e) = W_*(e) \times \frac{W_C(E)}{|T|} \)

\[
\text{EXCEPT}(C, e) \equiv \begin{cases} e \in M_{C \rightarrow G}(e) \\ |M_{C \rightarrow T}(C, M_{G \rightarrow T})| > \text{min\_weight} \\ X^2(C, e) > \chi^2_{0.05} \\ \text{WRAcc}(C, e) > 0 \end{cases}
\]

The contexts are assumed to be independent.

The total number of trips

The total number of trips in \( C \)

The total number of trips in \( e \)
**Exceptional contextual graph mining problem**

**Problem 4.1 (The Exceptional Contextual Graph Mining Problem).** Extracting meaningful patterns from an augmented graph $G = (V, E, T, \text{EDGE})$ is achieved by computing the theory:

$$
\{(C, CC_C) \mid CC_C = (V_{CC}, E_{CC}) \text{ is a maximal connected components of } G_C
\text{ with } G_C = (V, \{e \in E \mid \text{EXCEPT}(C, e) \text{ is true}\})
\text{ and } C \text{ is closed}
\text{ and } |V_{CC}| \geq \text{min\_vertex\_size}
\text{ and } |E_{CC}| \geq \text{min\_edge\_size}
\text{ and } \sum_{e \in E_{CC}} (\text{WRAcc}(C, e)) \geq \text{min\_sum\_wraacc}
\}$$

M. Plantevit, Pattern Mining in Augmented Graphs
Examples on Velo’v network

(gender = M, age = [14,26], univ_in, univ_out)

(zip_code = 38, gender = M, age = [26,60], pass = oura)
Including domain knowledge

\[ WRAcc(C, e) = \frac{1}{|T|} \left( W_c(e) - \bar{W}_c(e) \right) \]

\( \bar{W}_c(e) \) is the expected weight:

previously: \( \bar{W}_c(e) = W_*(e) \times \frac{W_c(E)}{|T|} \)

\( m(e) \) refer to the gravity model \( g(e) \) or the radiation model \( r(e) \).

The gravity model:

\[ g(e_{ij}) = n_i \times n_j \times f(d_{ij}) \]

\( n_i \) and \( n_j \) are respectively the populations of \( v_i \) and \( v_j \)

\( d_{ij} \) is the distance between \( v_i \) and \( v_j \), and

\( f(d_{ij}) \) represents the influence of the distance.
Expected weights are impacted

Example: a station located in Part Dieu
Examples

Only by mobility models

Context: * (all)

By the three models

Context: type = tecely, age ∈ [26,60]
Step back

- Need to provide instant results: output space sampling method.
- How to better present the results to user?
Exceptional subgraphs in vertex attributed graphs

Bendimerad et al., ICDM’16, KAIS 2018, MLG’18
Moranges et al., DS’18
Aims

- Discover subgraphs whose vertices have exceptional values on some attributes.
- Identify meaningful neighbourhood and describe them with their characteristics.

Foursquare venues

- Which model to capture exceptionality?
- Which quality measure?
- Which pattern language?
- How to extract the patterns?
Capturing exceptionality

**Characteristic**

A **characteristic** is \( S = (S^+, S^-) \) where \( S^+ \) and \( S^- \) two disjoint subsets of \( A \).

- \( S^+ \): positive trends.
- \( S^- \): negative trends.

**Example:**

- \( U = \{v_2, v_4\} \)
- \( S = (S^+ = \{Health\}, S^- = \{Tourism\}) \)

But how can we measure the relevance of \( S \) for \( U \)?

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

**Expected values**

<table>
<thead>
<tr>
<th>( v_1 )</th>
<th>( v_2 )</th>
<th>( v_3 )</th>
<th>( v_4 )</th>
<th>( v_5 )</th>
<th>( v_6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health</td>
<td>1</td>
<td>Health</td>
<td>9</td>
<td>Health</td>
<td>2</td>
</tr>
<tr>
<td>Tourism</td>
<td>7</td>
<td>Tourism</td>
<td>1</td>
<td>Tourism</td>
<td>6</td>
</tr>
<tr>
<td>Store</td>
<td>10</td>
<td>Store</td>
<td>9</td>
<td>Store</td>
<td>9</td>
</tr>
<tr>
<td>Food</td>
<td>4</td>
<td>Food</td>
<td>4</td>
<td>Food</td>
<td>4</td>
</tr>
</tbody>
</table>

| | 4.2 | 4.8 | 9.6 | 4.29 |
Exceptional subgraph mining problem

Given a graph \( G = (V, E, A) \), and two thresholds \( \sigma \) and \( \delta \), discover all exceptional sub-graphs \((U, S)\) such that:

1. \( |U| \geq \sigma \)
2. \( G[U] \) is connected
3. \( WRAcc(U, S) \geq \delta \)

Problem variants

- Closed exceptional subgraph to address redundancy issues
- Provide a sample of the output
- Output space sampling

CENERGETICS algorithm which fails with hundreds of attributes.

EXCESS algorithm
Examples

To describe and analyse cities

To understand the olfactory percept

Professional+, shop-
What about the user?

She can provide feedback about the patterns:
- How to take benefit from this feedback?
- Without changing the algorithm?

She has some priors about the data:
- How to model these priors?
- Use them to find really interesting patterns (according to her priors).

Taking into account user feedback into biased quality measures.
✓ Application to geolocated event detection on Twitter.

✓ Mining subjectively interesting attributed subgraphs
✓ MaxEntroy model to assess the interest of pattern.
✓ Trade-off between information content and pattern assimilation.
✓ Updating the model.
Unified framework for data-driven and user driven geolocated event discovery

\[ M_u(P) = \sum_{h \in H_p} \sum_{v \in K} \sum_{t \in I} \text{score}(h, v, t) \times \left( \frac{Q_h(h) + Q_v(v)}{2} \right) \]

\( Q_h : H \rightarrow [1, \text{maxPref}] \) and \( Q_v : V \rightarrow [1, \text{maxPref}] \) expresses respectively the interest of the term \( h \) and the vertex \( v \) to the user (and \( \text{maxPref} > 1 \)).

\[ Q_v(v) = \alpha \sum_{v' : (v, v') \in E} \frac{1}{\text{deg}(v)} \times Q_v(v') + (1 - \alpha) \times B(v) \]

\[ Q_h(h) = \alpha \sum_{h'} \frac{w(h, h')}{\text{deg}(h)} \times Q_h(h') + (1 - \alpha) \times B(h) \]

M. Plantevit, Pattern Mining in Augmented Graphs
The data-driven method outperforms state-of-the-art methods.

The user-driven method was assessed through an evaluation from the crowd (crowd flower).
Subjectively interesting subgraphs

The user may have some priors about the data.

"For each type, I know the number of places."

"For each area, I know the number of places."

User has to assimilate the patterns

"Please, give me patterns interesting and easy to assimilate."

The **subjective interestingness** as a trade-off between IC and DL:

\[
SI(U,S) = \frac{IC(U,S)}{DL(U,S)}
\]

**SIAS-Miner algorithm**

**Equality Constraints**

**Background Knowledge**

**Information Content**

\[
IC(U,S) = -\log(Pr(U,S))
\]

**Assimilation cost**

**Description Length**

\[
DL(U,S) = DL_A(S) + DL_V(U)
\]

The vertices that are at a distance of at most 2 of the vertex A and the vertex B.

M. Plantevit, Pattern Mining in Augmented Graphs
SIAS Miner: Examples

<table>
<thead>
<tr>
<th>ID</th>
<th>Exceptionally prevalent</th>
<th>Exceptionally non-prevalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>nightlife, professional</td>
<td>college, event</td>
</tr>
<tr>
<td>$P_2$</td>
<td>nightlife, food</td>
<td>college</td>
</tr>
<tr>
<td>$P_3$</td>
<td>food</td>
<td></td>
</tr>
<tr>
<td>$P_4$</td>
<td>food</td>
<td>college</td>
</tr>
<tr>
<td>$P_{14}$</td>
<td>professional</td>
<td></td>
</tr>
</tbody>
</table>
Step back: have we filled the gap between the user and her data?

Not yet, but we are in the good direction!

✓ Complexity of the data
✓ Complexity of the domain
✓ Complexity of the user
  ✓ Background knowledge
  ✓ User feedback
✓ Complexity of the output

- Incorporating of non-ordinal attribute types
- Integrating other kinds of prior beliefs (e.g., correlation)
Conclusion and Future Directions

M. Plantevit, Pattern Mining in Augmented Graphs
Conclusion: take away message

✅ Augmented graphs as a powerful way to model real-world phenomena.

⚠ Take care of all the complexities:
  • data, user, domain and output
This is the only mean to provide actionable insights and boost human knowledge.
Is there a future for pattern mining?

**YES ... if we**
- Democratize the pattern mining tools (Knime, Weka, Python libraries)
- Make them easily usable

Analysts will always need of descriptive analysis techniques:
- Crystal clear descriptive solutions

Still some issues to tackle!
Improve pattern rendering

Data Mining meets Visualization and Information Retrieval.

• Highlight the interest of the pattern with respect to:
  • the user priors,
  • the domain knowledge,
  • how this pattern stands out from the others.

⚠ Decades of investigation done in the Visualization research community.

• Same observation with the Information Retrieval community.
  • (re-)ranking results, taking into account user interest and her satisfaction, recommending are at the heart of the studies from this community.
Causality

- Users tend to assimilate patterns/rules with a causality point of view.

  Enthusiasm is quickly followed by disappointment.

- Pattern mining will bring as much as hope as disappointment as long as the problem of causality is not solved.

- A timely challenge!
Pattern mining as the corner stone of data science projects

- Pattern mining to foster **interdisciplinarity**.
- The pattern syntax can be easily understood by any scientist.
- Domain knowledge of each discipline can be integrated.
- Discovered patterns as an excellent support for discussion between the scientists from different disciplines.
- **ROI**: such projects also provide new challenges in data mining.
Describe as simplest as possible!

Towards multiple-pattern domains pattern set mining.

---

**Strong assumption**: the user knows the good pattern domain.

⚠ A too complex pattern domain: over-described results and misleading interpretation.

⚠ A too simple pattern domain: impossible to describe some complex phenomena.

- Automatically find the good level of description (pattern syntax).
  - Itemsets (w/ | w/o) numerical values, sequence, graphs, 3d graphs, ...
Data Mining meets Machine Learning: Towards sparse and interpretable Deep Neural Networks

- Obvious need of effective predictive models
- Work into the models to understand / simplify them
- Work on the I/O to both understand and improve the models.
- Investigate languages to characterize them.
THE END

THE BEGINNING