Mining Subjectively Interesting Attributed Subgraphs

Ahmad Mel <ahmad.mel@ugent.be>

GHENT UNIVERSITY

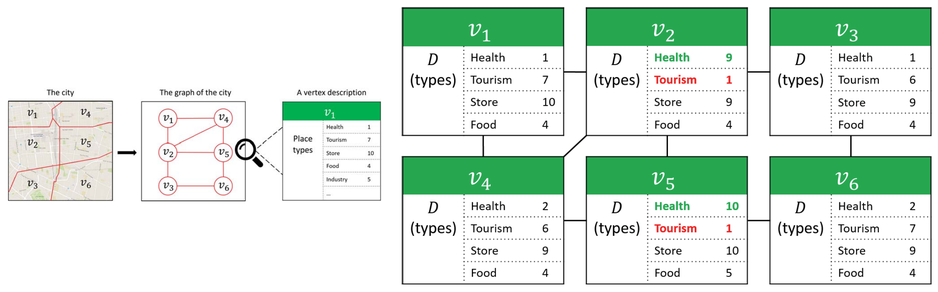
IDLab INTERNET & DATA LAB

Introduction

- **Graph mining**:
  Extracting meaningful relationships from connected data objects (vertices).
- **Community Detection**:
  Identification of informative subgraphs, i.e. groups of vertices that belong together.
- **Task**:
  Finding interesting subgraphs, i.e. cohesive subgraphs whose attribute values are similar within the subgraph but exceptional when compared to the rest of the graph.

Data

- **Foursquare Geographical Graph:** Each vertex depicts a district in a city and edges link adjacent districts. Each attribute represents the number of venues of a sepecific type.



- **Ingredients Graph:** Each vertex is a recipe ingredient, while the attributes correspond to the number of occurences of the ingredient in the recipe of each type of cuisine. An edge exists between two ingredients if their Jaccards similarity between their recipies is higher than a certain threshold. It's a big ugly graph.

Pattern Language

Given a graph $G = (V, E, \hat{A})$

V is a set of n vertices

$E \subseteq V \times V$ is a set of m edges

$\hat{A}$ is a set of p numerical attributes on vertices

$\hat{a}(v) \in \text{Dom}_a$ is the value of attribute $\hat{a} \in \hat{A}$ on $v \in V$.

$N_d(v) = \{u \in V \mid dist(v, u) \leq d\}$.

$\mathcal{N} = \{N_d(v) \mid v \in V \wedge d \in 0, D\}$

We define a CSEA pattern (Cohesive Subgraphs with Exceptional Attributes) as a tuple (U, S) such that

- $U \subseteq V$
- $S \subseteq \{(a, [k_a, \ell_a]) \mid a \in A\}$
- $\forall (a, [k_a, \ell_a]) \in S$
- $\forall u \in U,\ k_a \leq \hat{a}(u) \leq \ell_a$.

Subjective Interestingness of CSEA patterns

$\text{SI}(U, S) = \frac{\text{IC}(U,S)}{\text{DL}(U,S)}$.

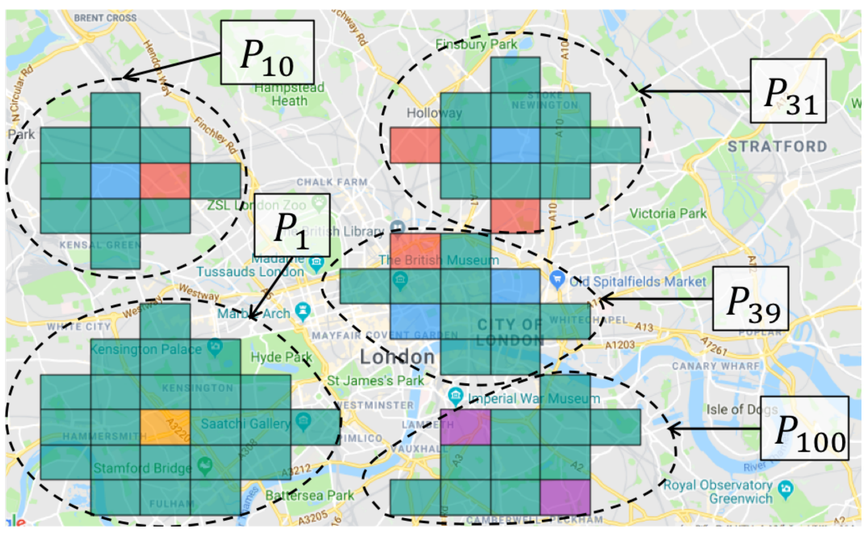
- $\text{IC}(U, S) = -\log(\text{Pr}(U, S))$.
  IC is the amount of information provided by showing the user a pattern. The quantification is based on the gain from a Maximum Entropy background model that delineates the current knowledge of a user.
  $\text{Pr}(c_a(v) \in [k_{c_a}, \ell_{c_a}]) = \ell_{c_a} - k_{c_a} + p_{av}k_{c_a}$.
  $\text{IC}(U, S) = -\sum_{(a,[k_{c_a},\ell_{c_a}])\in S} \sum_{v \in U} \log(\ell_{c_a} - k_{c_a} + p_{av}k_{c_a})$.
- $\text{DL}(U, S) = \text{DL}_A(S) + \text{DL}_V(U)$:
  The DL assesses the complexity of reading a pattern, the user being interested in concise and intuitive descriptions. Thus, we describe a set of vertices as an intersection of neighborhoods of certain distance from certain vertices, the distance and vertices making up the description of the subgraph. $\mathcal{N}(U) = \{N_d(v) \in \mathcal{N} \mid U \subseteq N_d(v)\}$
  $f : 2^{\mathcal{N}(U)} \times U \longrightarrow \mathbb{R}$:
  $f(X, U) = (|X| + 1) \cdot \log(|\mathcal{N}|) + (|(X, U)| + 1) \cdot \log(|\cap_{x \in X} x|)$
  $\text{DL}_V(U) = \min_{X \subseteq \mathcal{N}(U)} f(X, U)$.

SIAS-Miner

SIAS-Miner mines interesting patterns using an enumerate-and-rank approach.

- First, it enumerates all CSEA patterns (U, S) that are closed simultaneously with respect to U, S, and the neighborhood description $\mathcal{N}(U)$. Closing a set of vertices U w.r.t. S and $\mathcal{N}(U)$ would:
  + Maximize $\text{IC}(U)$, and minimizes $\text{DL}_V(U)$.
  - Increase the value of $\text{DL}_A(S)$,
  + Allow to drastically improve the performance of the algorithm
  + Reduce the size of the output, without altering the result quality
- Second, it ranks patterns according to their SI values.
  – The calculation of $\text{IC}(U, S)$ and $\text{DL}_A(S)$ is direct.
  – Computing $\text{DL}_V(U)$ is not trivial. We implement an efficient branch-and-bound algorithm, coupled with different pruning techniques to reduce the search space, that calculates the minimal description of U and stores the results.

London Results



Patterns	$S = \{(a, [\ell_{c_a}, k_{c_a}])\}$	$SI(P)$
P_1	$\{\text{food} : [0, 0.31]\}^+$	0.69
P_{10}	$\{\text{food} : [0, 0.46]\}^+, \{\text{art} : [0.57, 1], \text{college} : [1, 1], \text{event} : [1, 1]\}^-$	0.54
P_{31}	$\{\text{nightlife} : [0, 0.44], \text{food} : [0, 0.31]\}^+$	0.49
P_{39}	$\{\text{professional} : [0, 0.40], \text{college}[0, 0.46], \text{outdoors}[0, 0.47]\}^+$	0.48
P_{100}	$\{\text{food} : [0, 0.46]\}^+, \{\text{college} : [1, 1], \text{event} : [1, 1]\}^-$	0.43

Green cells represents vertices covered by a CSEA pattern

Blue cells are the centers

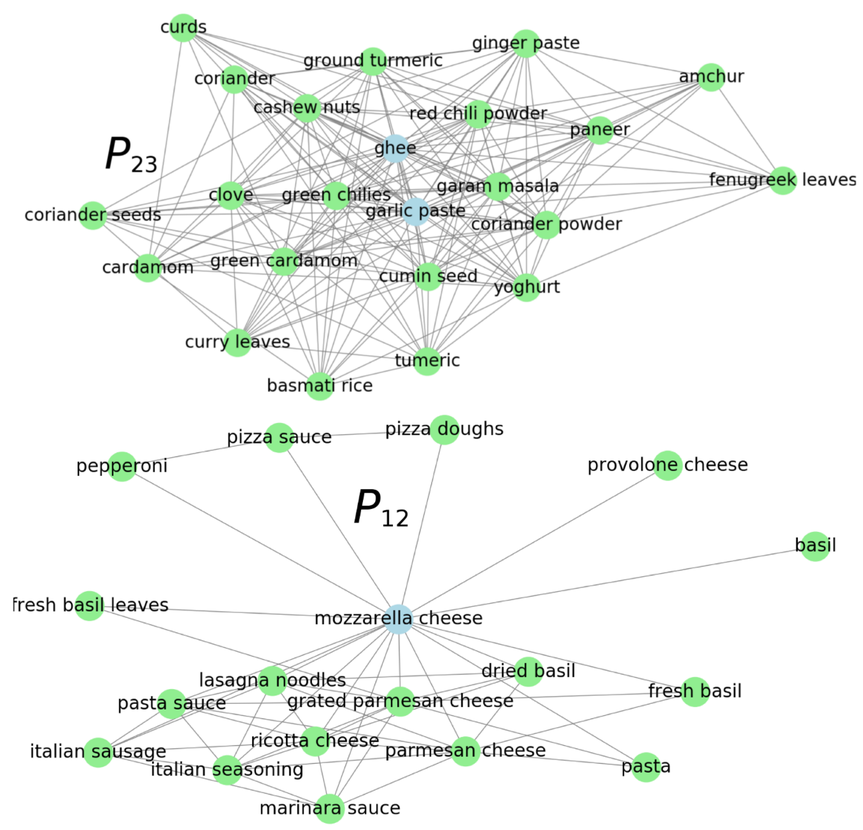
Purples cells are the centers that do not belong to the pattern,

Orange cells are centers that are also exception (i.e. behave differently from the pattern but covered by the description)

Red cells are normal exceptions.

- P_1 covers the neighborhood of range 3 from the orange vertex, with an overexpression of Food venues.
- P_{39} is described by the intersection of the neighbors of the blue vertices with a maximum distance of 3. It covers the City of London where there is a significant amount of professional venues, college, universities, and outdoor venues.

Ingredients Results



Patterns	$S = \{(a, [\ell_{c_a}, k_{c_a}])\}$	$SI(P)$
P_{12}	$\{\text{Italian} : [0, 10^{-18}]\}^+$	17.93
P_{10}	$\{\text{Indian} : [0, 10^{-18}], \text{Japanese} : [0, 0.44]\}^+$	0.54

Above are two patterns discovered by SIAS-Miner in the Ingredients graph.

- P_{12} corresponds to a set of ingredients that appear a lot in Italian recipes. They are described as neighbors of mozarella cheese, with two exceptions.
- P_{23} consists in some ingredients that are over expressed on Indian and Japanese recipes. They can be expressed as the neighbors of both ghee and garlic paste, with 6 exceptions.

Conclusion

- We introduced a new pattern syntax in attributed graphs, CSEA, that, given a vertex attribtued graph, would provide the user with a set of attributes that have exceptional values throughout a subset of vertices.
- The defnition of interestingness is based on information theory, as the ratio of the information content (IC) over the description length DL.
- We have proposed an effective algorithm that enumerates and ranks patterns of this language.
- Empirical results on 2 real-world datasets confirm that CSEA patterns are intuitive, and the interestingness measure aligns well with actual subjective interestingness.

References

1. SIAS-Miner: Mining Subjectively Interesting Attributed Subgraphs - Anes Bendimerad, Ahmad Mel, Jefrey Lijffijt, Marc Plantevit, Céline Robardet, Tijl De Bie
2. Unsupervised exceptional attributed sub-graph mining in urban data - Anes Bendimerad, Marc Plantevit, Céline Robardet
3. Maximum entropy models and subjective interestingness - Tijl De Bie

Funded by: The European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Grant Agreement no. 615517.

Presented at the DS³ Data Science Summer School in Palaiseau on 2018-07-25