Exploring Multiple Clustering in Attributed Graphs

Gustavo Paiva Guedes[1,2], Eduardo Ogasawara[2],
Eduardo Bezerra[2], Geraldo Xexéo[1]
Federal University of Rio de Janeiro (COPPE/UFRJ)
Federal Center of Technological Education of Rio de Janeiro (CEFET/RJ)

Introduction

- The goal of graph clustering is to cluster vertices of a graph, in such a way that vertices inside a cluster are densely connected, and vertices of different clusters are sparsely connected.
- Most clustering algorithms identify only one partition of the data [1]. However a single partition may not provide sufficient insight.
- Several areas, such as marketing in social networks, demands the exploration of multiple clustering solutions. Thus, it would be interesting to explore other clustering solutions.

Multiple clustering paradigm

- The main goal of multiple clustering algorithms is to produce multiple clustering solutions from the same collection of objects that can reveal novel ways of interpreting the same data.
- In this work, we provide multiple clustering solutions by combining vertex attributes with graph structure. This type of graph is called attributed graphs [2].

Problem Statement

- An attributed graph G is a 4-tuple $G = (V, E, \Lambda)$, where $V = \{v_1, v_2, \ldots, v_n\}$ is a set of n vertices, E is a set of edges, which are unordered pairs of elements of V, $\Lambda = \{a_1, a_2, \ldots, a_m\}$ is a set of m attributes. In an attributed graph G, each vertex v_i is associated with an attribute vector of length m.
- Given an attributed graph $G = (V, E, \Lambda)$, we tackle the problem of generating a set of alternative non-redundant clusterings of vertices by combining both topological and relational information.

Combining structure with vertex attributes

- Given a set of attributes, CRAG algorithm generates one clustering solution that combines graph structure with vertex attribute information.

Algorithm CRAG$(G, attrSet, d, k)$

Input:
- Attributed Graph $G = (V, E, \Lambda, f)$
- $attrSet$ = set of attributes
- d = neighborhood distance
- k = number of clusters

Output: one clustering solution of vertices in G.

$E' \leftarrow \emptyset$
$s \leftarrow \text{getSimilarityThreshold}(G, attrSet, d)$
for all $v_i \in G.V$ **do**
  $\mathcal{N}_{v_i} \leftarrow \text{getNeighborhood}(G, v_i, d)$
  for all $v_j \in \mathcal{N}_{v_i}$ **do**
    if $\text{similarity}(v_i, v_j, attrSet) \geq s$ **then**
      $E' \leftarrow E' \cup \{edge(v_i, v_j)\}$
    end if
  end for
end for
$G.E \leftarrow G.E \cup E'$
return $\text{spectralClustering}(G, k)$

Comparison of clusterings

- We use density (D), entropy (S) and NMI to evaluate quality of clusterings. Density represents the sum of the number of intra-group edges divided by the total number of edges. Density value is a number between 0 and 1. Entropy is an uncertainty measure for a probability distribution. Entropy tends to decrease if distribution inside groups are uniform. We normalized entropy to represent a number between 0 and 1. NMI is used to compere two clusterings. It has also values between 0 and 1. If two clusterings are similar, NMI tends to increase.

Multiple clustering solutions

- The aim of M-CRAG algorithm is to generate multiple clustering solutions using CRAG algorithm.

Algorithm M-CRAG(G, d, t, k)

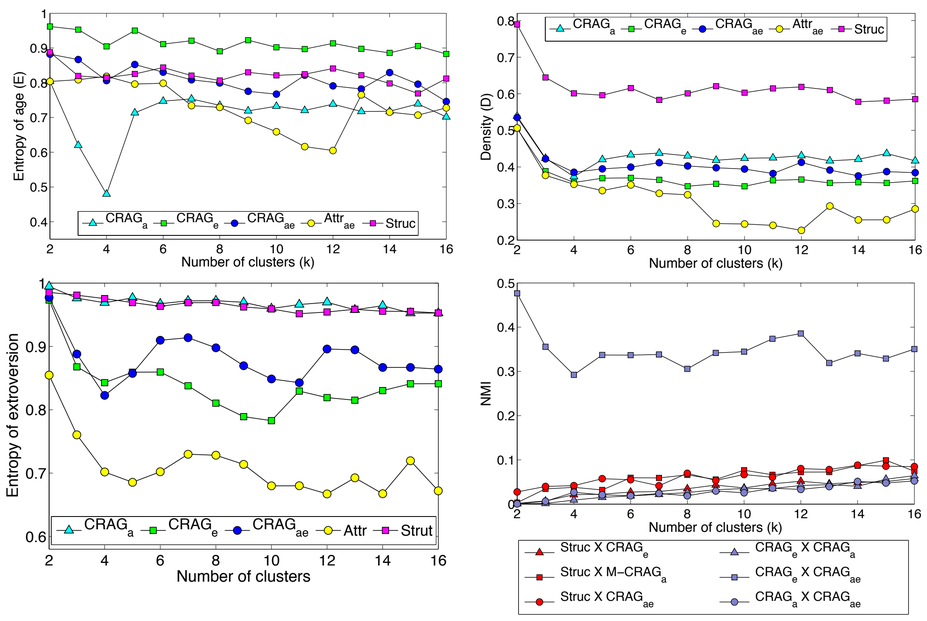
Input:
- Attributed Graph $G = (V, E, \Lambda, f)$
- d = neighborhood distance
- t = Maximum threshold for NMI
- k = number of clusters

Output: $\mathcal{C}$, a set of non-redundant clustering solutions of vertices in G.

$\mathcal{C} \leftarrow \emptyset$
$\mathcal{F} = \text{chooseAttributes}(G.\Lambda)$
for all $attrSet \in \mathcal{F}$ **do**
  $c \leftarrow \text{CRAG}(G, attrSet, d, k)$
  if $\text{MAX}_{\text{NMI}}(\mathcal{C}, c) \leq t$ **then**
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{c\}$
  end if
end for
return $\mathcal{C}$

Experimental Analysis

- Dataset. We used MQD500 [4] which is an attributed graph dataset extracted from MQD Brazilian online social network [3] .
- Experimental Settings. We compared three kinds of clusterings for each dataset: (i) *Struc* refers to clusterings generated using spectral clustering that uses only graph structure. (ii) Attr_l refers to clusterings generated using only vertex attributes. (iii) CRAG_l refers to clusterings produced using M-GRAG.



Conclusions

- We explored multiple clustering solutions in attributed graphs using M-CRAG algorithm. Experimental results using MQD500 dataset showed that M-CRAG is able to produce non-redundant alternative clustering solutions with good quality.

References

[1] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review, 1999.
[2] Y. Zhou, H. Cheng, and J. X. Yu. Graph clustering based on structural/attribute similarities. Proc. VLDB Endow., 2(1):718{729, Aug. 2009.
[3] Meu querido diario. http://www.meuqueridodiario.com.br.
[4] Mqd500 dataset. http://sourceforge.net/p/gpca/wiki/MQD500/.

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico
FAPERJ Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro