Comparing Graph Partitions

Valérie Poulin
François Théberge*
theberge@ieee.org

August 2019

Outline

1. Introduction: graph clustering
2. Common similarity measures
3. Link to binary classification
4. Graph-aware measures
5. Topological features

Graph Clustering

$G = (V, E)$, $E \subset V \times V$, $|V| = n$, $|E| = m$

A, adjacency matrix: $a_{ij} = 1 \Leftrightarrow (i, j) \in E$

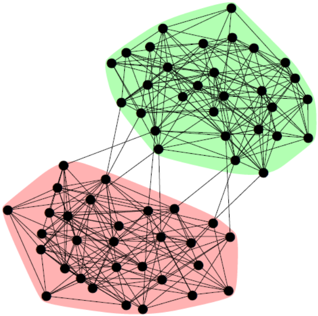
d_i: degree of vertex i.

- **Graph clustering/partitioning**
  Partition the vertices into connected subgraphs
- **Community finding**
  Not all vertices need to be assigned to a cluster
- **Fuzzy clustering**
  Vertices are members of no, one or many clusters.

Graph Clustering

Graph partition: $\mathbf{A} = \{A_1, A_2, ..., A_k\}$, partition of V

- Each A_i induces a connected subgraph
- Generalization of connected components
  - *Large density of edges within clusters*
  - *Low density of edges between clusters*



Graph Clustering

Graph clustering is an important tool for relational EDA:

- Graph size reduction
- Community detection
- Anomaly detection, etc.

How to pick a clustering algorithm?

- Quality of the clusters
- Stability
- Efficiency (time and space)
- Other nice features:
  - No need to specify the number of clusters (k).
  - Hierarchy of clusters.

Graph Clustering

This is **unsupervised** learning.
There is **no** clear objective function for graph clustering.
Algorithms use different functions such as:

- Modularity:

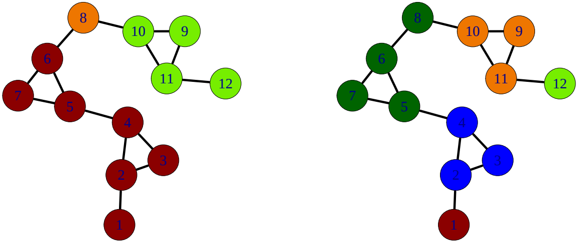
$$Q = \frac{1}{2m} \sum_{\substack{i,j \text{ same} \\ \text{cluster}}} \left(a_{ij} - \frac{d_i d_j}{2m} \right)$$

- Ncut:

$$\sum_i \frac{cut\left(A_i, \overline{A_i}\right)}{\#\text{ edges in } A_i}$$

Compare different partitions

- Measure of quality: $sim(\mathbf{T}, \mathbf{A})$ w.r.t. ground truth partition $\mathbf{T}$
- Measure of stability: $sim(\mathbf{A}, \mathbf{A}')$ for several runs of the same algorithm
- Compare results between algorithms: $sim(\mathbf{A}, \mathbf{B})$.



Families of measures

Several measures have the following format:

- Pairwise-counting based:

$$PW_f(\mathbf{A}, \mathbf{B}) = \frac{|P_A \cap P_B|}{f(|P_A|, |P_B|)}$$

- Information-based:

$$MI_f(\mathbf{A}, \mathbf{B}) = \frac{I(\mathbf{A}, \mathbf{B})}{f(H(\mathbf{A}), H(\mathbf{B}))}$$

- χ^2-based ($|A| = k$ and $|B| = r$):

$$X^2_f(\mathbf{A}, \mathbf{B}) = \frac{X^2(\mathbf{A}, \mathbf{B})}{f((k-1), (r-1))}$$

where $f(x, y) \in \{\min(x, y), \max(x, y), \text{mean}(x, y), \sqrt{xy}\}$.

Pairwise-based family

Let:
$\mathbf{A} = (A_1, ..., A_k) = (\{1, 2, .., 7\}, \{8\}, \cdots)$
$\mathbf{B} = (B_1, ..., B_r) = (\{1\}, \{2, 3, 4\}, \{5, 6, 7, 8\}, \cdots)$
Those measures are based on pairs of elements clustered together in both A and in B:

$$\begin{aligned} P_A &= \{(1,2),(1,3),(1,4),(1,6),(1,7),(2,3),\cdots\} \\ P_B &= \{(2,3),(2,4),(3,4),(5,6),(5,7),(5,8),\cdots\} \end{aligned}$$

Key quantity is $|P_A \cap P_B|$.

Pairwise-based family

Examples include the Jaccard index:

$$\frac{|P_A \cap P_B|}{|P_A \cup P_B|}$$

and the RAND index:

$$\frac{|P_A \cap P_B| + |\overline{P_A} \cap \overline{P_B}|}{\binom{n}{2}}$$

Information-based family

Based on the mutual information between A and B.

Key quantity is:

$$I(\mathbf{A}, \mathbf{B}) = \sum_{i,j} \frac{|A_i \cap B_j|}{n} \log \frac{|A_i \cap B_j|/n}{|A_i||B_j|/n^2}$$

Example: Normalized mutual information (NMI):

$$\frac{I(\mathbf{A}, \mathbf{B})}{(H(\mathbf{A}) + H(\mathbf{B}))/2}$$

χ^2-based family

Key quantity is:

$$X^2(\mathbf{A}, \mathbf{B}) = \sum_{i,j} \frac{1}{|A_i||B_j|} \left(|A_i \cap B_j| - \frac{|A_i||B_j|}{n} \right)^2$$

Examples: Cramer's V and Tschurprow's T measures.

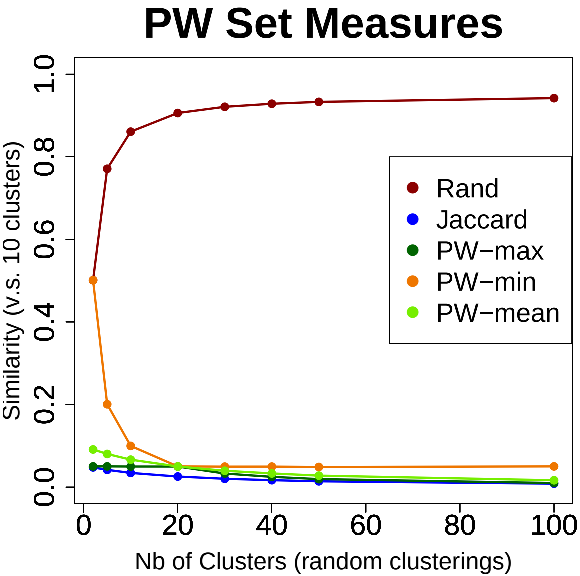
Measures vs. size distribution

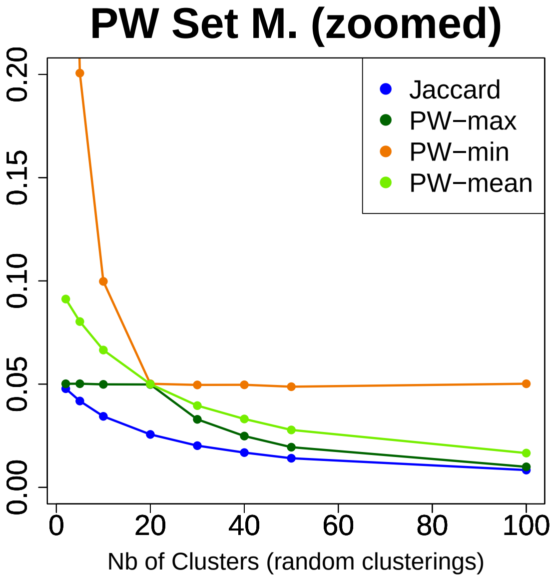
Q: *How do the measures behave when comparing partitions of different sizes?*

Experiment (repeated many times):

- $\mathbf{A}$, a partition of V with $|\mathbf{V}| = 10$
- $\mathbf{B}^{(t)}$, random partitions of V with $|\mathbf{B}^{(t)}| = t$,
  for $t = 2, 5, 10, 20, 30, 40, 50$ and 100.
- Measure similarity between $\mathbf{A}$ and all partitions $\mathbf{B}^{(t)}$.

...and hope all similarity values are VERY low!


PW Set Measures
Similarity (v.s. 10 clusters)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Nb of Clusters (random clusterings)
Rand
Jaccard
PW−max
PW−min
PW−mean


PW Set M. (zoomed)
Jaccard
PW−max
PW−min
PW−mean
0.00
0.05
0.10
0.15
0.20
0
20
40
60
80
100
Nb of Clusters (random clusterings)

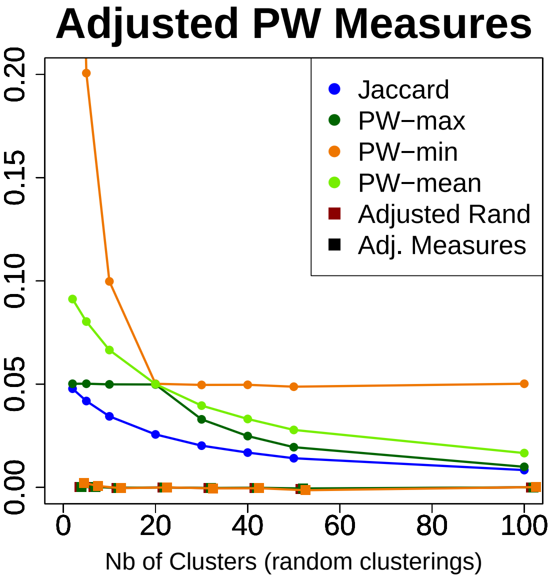
Adjustment for chance

$$\text{Adjusted Similarity}(\mathbf{A},\mathbf{B}) = \frac{\text{Similarity}(\mathbf{A},\mathbf{B}) - \text{Expected Sim}(|A_i|'s,|B_j|'s)}{1 - \text{Expected Sim}(|A_i|'s,|B_j|'s)}$$

- Adjusted Forms for the pairwise measures:

$$APW_f(\mathbf{A},\mathbf{B}) = \frac{|P_A \cap P_B| - |P_A||P_B|/\binom{n}{2}}{f(|P_A|,|P_B|) - |P_A||P_B|/\binom{n}{2}}$$

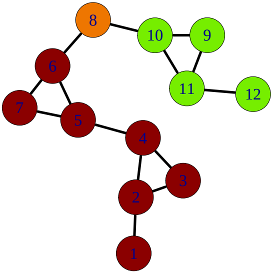
- Jaccard has no known adjusted form.
- $\text{ARI}(\mathbf{A},\mathbf{B}) = APW_{mean}(\mathbf{A},\mathbf{B})$, the **adjusted RAND index.**

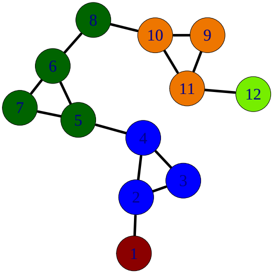

Adjusted PW Measures
Jaccard
PW−max
PW−min
PW−mean
Adjusted Rand
Adj. Measures
0.00
0.05
0.10
0.15
0.20
0
20
40
60
80
100
Nb of Clusters (random clusterings)

Adjustment for chance

- Information theoretic-based and χ^2-based can also be adjusted for chance
- We mostly use the following two measures:
  - ARI - adjusted RAND index
  - AMI - adjusted mutual information

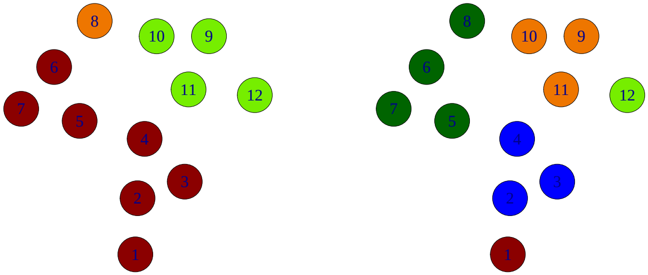
Similarity Measures between graph partitions





We have adjusted measures to compare partitions. ✓

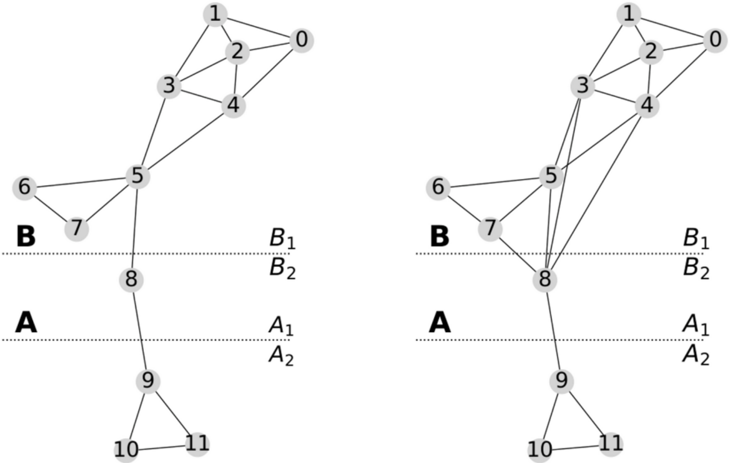
Similarity Measures between graph partitions



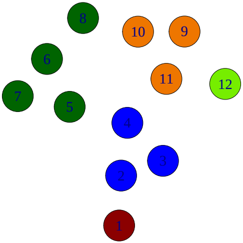
... but we do not consider the graph topology at all!

Should edges be considered when measuring similarity?

Graph-aware measures?



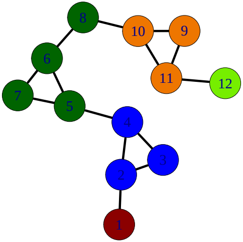
Edge classification



A graph partition can be represented by set partition on V.
In the graph above,

$$\mathbf{A} = (\{1\}, \{2, 3, 4\}, \{5, 6, 7, 8\}, \{9, 10, 11\}, \{12\}).$$

Edge classification



We can also consider binary edge classification (with vertices in same cluster or not). In the graph above:

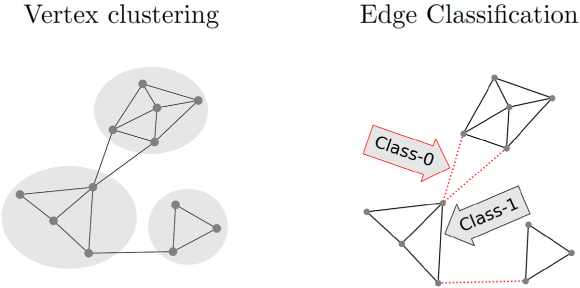
$$(2,3),(2,4),(3,4),...,(9,10),(9,11),(10,11) \rightarrow \text{class } 1$$

$$(1,2),(4,5),(8,10),(11,12) \rightarrow \text{class } 0.$$

Edge classification

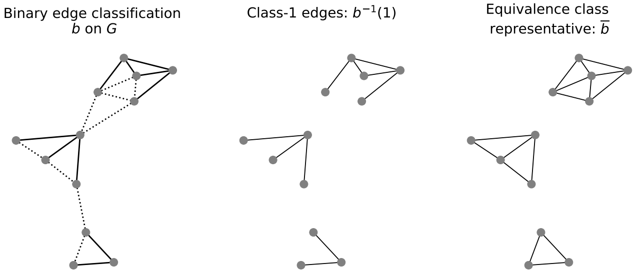
More formally, for a vertex partition $\mathbf{A}$ we define the binary vector $b_{\mathbf{A}}$ of length m where, for each edge $e = (i, j) \in E$:

$$b_{\mathbf{A}}(e) = \begin{cases} 1 & \exists A_k \in \mathbf{A} \mid i, j \in A_k \\ 0 & \text{otherwise.} \end{cases}$$



Edge classification

And we can go the other way around (modulo the fact that some edges may be de-facto mapped to class 1):



Evaluation of binary classifiers

Consider b_A and b_B, two binary edge classifiers.
The **four base counts** for comparing binary classifiers are:

b_A/b_B	1	0
1	$\lvert P_A \cap P_B \cap E \rvert$	$\lvert P_A \cap \overline{P_B} \cap E \rvert$
0	$\lvert \overline{P_A} \cap P_B \cap E \rvert$	$\lvert \overline{P_A} \cap \overline{P_B} \cap E \rvert$

Graph-aware clustering measures

Accuracy	gR: $\frac{\vert P_A \cap P_B \cap E\vert + \vert\overline{P_A} \cap \overline{P_B} \cap E\vert}{\vert E\vert}$
Jaccard	gJ: $\frac{\vert P_A \cap P_B \cap E\vert}{\vert(P_A \cup P_B) \cap E\vert}$
F-score ($\beta = 1$)	gPW_{mn}: $\frac{\vert P_A \cap P_B \cap E\vert}{\frac{1}{2}(\vert P_A \cap E\vert + \vert P_B \cap E\vert)}$
Cosine	gPW_{gmn}: $\frac{\vert P_A \cap P_B \cap E\vert}{\sqrt{\vert P_A \cap E\vert\vert P_B \cap E\vert}}$
Simpson	gPW_{min}: $\frac{\vert P_A \cap P_B \cap E\vert}{\min\{\vert P_A \cap E\vert, \vert P_B \cap E\vert\}}$
Braun&Banquet	gPW_{max}: $\frac{\vert P_A \cap P_B \cap E\vert}{\max\{\vert P_A \cap E\vert, \vert P_B \cap E\vert\}}$

Adjusting the graph-aware measures

From the binary edge vectors:

$$|P_{\mathbf{A}} \cap P_{\mathbf{B}} \cap E| = |b_{\mathbf{A}} \cdot b_{\mathbf{B}}|$$

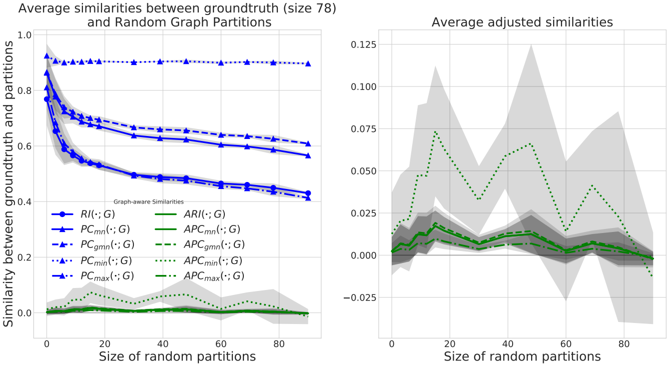
we can write a family of pairwise-counting graph-aware measures:

$$PC_f(\mathbf{A},\mathbf{B};G) = \frac{|b_{\mathbf{A}} \cdot b_{\mathbf{B}}|}{f(|b_{\mathbf{A}}|,|b_{\mathbf{B}}|)}, \quad APC_f(\mathbf{A},\mathbf{B};G) = \frac{|b_{\mathbf{A}} \cdot b_{\mathbf{B}}| - \frac{|b_{\mathbf{A}}|\cdot|b_{\mathbf{B}}|}{|E|}}{f(|b_{\mathbf{A}}|,|b_{\mathbf{B}}|) - \frac{|b_{\mathbf{A}}|\cdot|b_{\mathbf{B}}|}{|E|}}$$

where we used a naive adjustment on the right-hand side.

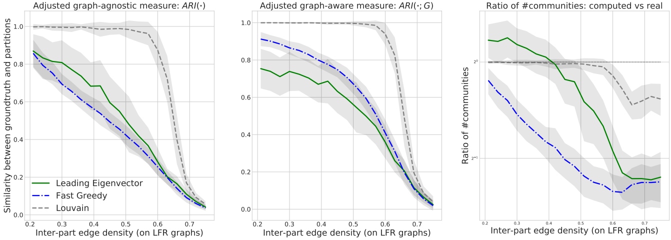
Adjusting the graph-aware measures

Results on LFR graphs:



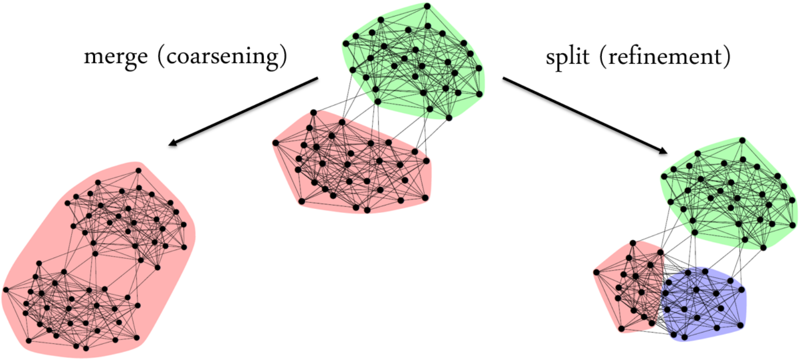
Graph-aware similarity measures

Conclusions may vary depending on the type of measure ...



Complementarity

Graph-aware and agnostic measures have opposite behaviours with respect to resolution issues.



Complementarity

Let G with ground-truth community $\mathbf{A}$ and let $\mathbf{B_1}$ and $\mathbf{B_2}$ be a coarsening and a refinement of $\mathbf{A}$ respectively.

Under some conditions, $\mathbf{A}$ is closer to $\mathbf{B_2}$ than to $\mathbf{B_1}$ under the graph-agnostic measures, and the opposite is true under the graph-aware measures.

Larger number of clusters are favoured when using graph-agnostic measures, smaller number for graph-aware measures.

Getting high values with respect to **both** measures is desirable.

Complementarity

Consider $\mathcal{G}(n, p, q, \mathbf{A})$, a variant of Girvan and Newman model to study a family of graphs having community structure.

Graphs in $\mathcal{G}(n, p, q, \mathbf{A})$ have n vertices split into a partition $\mathbf{A}$ with p (resp. q) the proportion of randomly selected pairs of vertices in *same* (resp. *different*) parts of $\mathbf{A}$ sharing an edge.

Complementarity

Graph-aware and agnostic measures penalize in opposite ways:

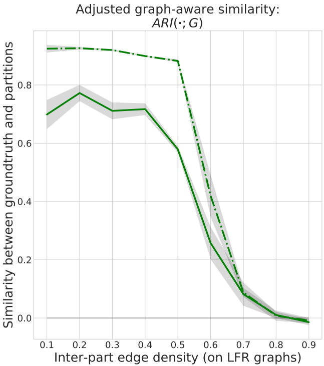
Theorem 1 *Consider* $G_{\mathbf{A}} \sim \mathscr{G}(n,p,q,\mathbf{A})$ *with* $\mathbf{B}_1 > \mathbf{A}$ *a coarsening of* $\mathbf{A}$ *and* $\mathbf{B}_2 < \mathbf{A}$, *a refinement of* $\mathbf{A}$ *such that* $|P_{\mathbf{A}}|^2 < |P_{\mathbf{B}_1}| \cdot |P_{\mathbf{B}_2}|$. *Then*

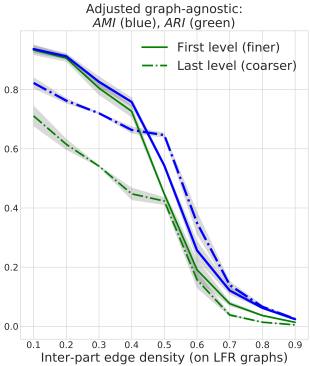
(i) $PC_{mn}(\mathbf{A},\mathbf{B}_1) < PC_{mn}(\mathbf{A},\mathbf{B}_2)$.

(ii) $\mathbb{E}_{G_{\mathbf{A}}}[PC_{mn}(\mathbf{A},\mathbf{B}_1;G_{\mathbf{A}})] > \mathbb{E}_{G_{\mathbf{A}}}[PC_{mn}(\mathbf{A},\mathbf{B}_2;G_{\mathbf{A}})]$, *if* $p > q\frac{|P_{\mathbf{B}_1} \setminus P_{\mathbf{A}}|}{|P_{\mathbf{A}} \setminus P_{\mathbf{B}_2}|}$.

Complementarity

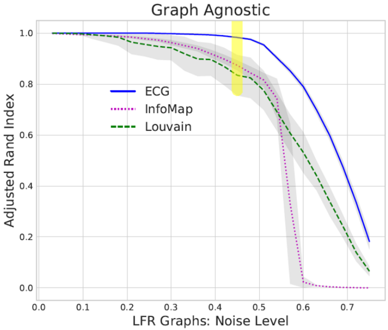
Comparing multilevel Louvain (first and last levels):





Complementarity

Sneak preview: ECG vs other state-of-the-art algorithms:



Code

Code available on CodeOcean:

COMPUTER SCIENCE *July 1, 2018*

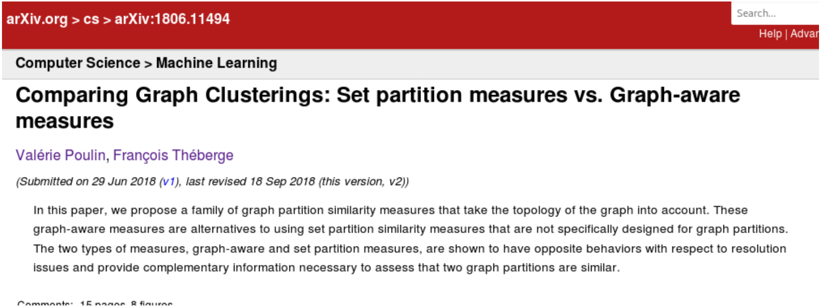
Adjusted graph-aware Rand Index for comparing graph partitions

Valérie Poulin, François Théberge

Adjusted graph-aware Rand Index for comparing graph partitions. We propose an adjusted graph partition similarity measure that take the topology of the graph into account. This graph-aware measure is an alternative to using set partition similarity measures that are not specifically designed for graph partitions. The two types of measures, graph-aware and set partition measures, are shown to have opposite behaviours with respect to resolution issues and provide complementary information necessary to assess...

Paper

Paper on arXiv:



arXiv.org > cs > arXiv:1806.11494

Search...

Help | Advar

Computer Science > Machine Learning

Comparing Graph Clusterings: Set partition measures vs. Graph-aware measures

Valérie Poulin, François Théberge

(Submitted on 29 Jun 2018 (v1), last revised 18 Sep 2018 (this version, v2))

In this paper, we propose a family of graph partition similarity measures that take the topology of the graph into account. These graph-aware measures are alternatives to using set partition similarity measures that are not specifically designed for graph partitions. The two types of measures, graph-aware and set partition measures, are shown to have opposite behaviors with respect to resolution issues and provide complementary information necessary to assess that two graph partitions are similar.

Comments: 15 pages, 8 figures

Topological features

Another way to validate clustering(s) is to compare topological features of the clusters.

Several measures are proposed in: Orman *et al.*, arXiv:1206.4987

Some examples are, for community c with n_c nodes and m_c edges:

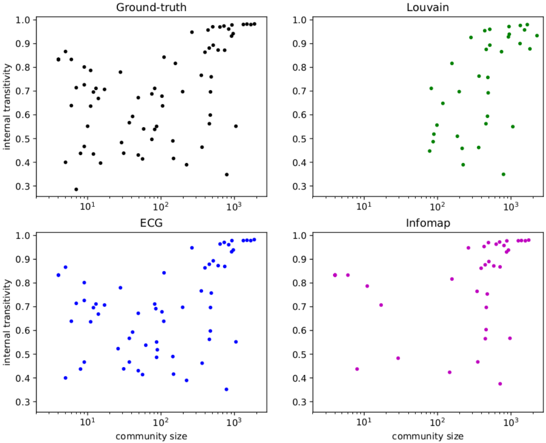
- **scaled density:** $n_c \cdot m_c / \binom{n_c}{2}$
- **internal transitivity:**

$$\frac{1}{n_c} \sum_{i \in c} \frac{e_c(i)}{\binom{d_c(i)}{2}}$$

where $e_c(i)$ is the number of edges between neighbours of i within c, and $d_c(i)$ in the degree of i within c.

Topological features

Features can then be compared as a function of the cluster sizes.



Conclusion

Take away:

- Use **adjusted** set-based similarity measures

  ... reduces the bias of measures on granularity of partitions.
- Graph-agnostic (ARI, AMI) and graph-aware (AGRI) measures are complementary:

  ... they **should be used simultaneously** when assessing the superiority of an algorithm.

Comparing Graph Partitions

Notebook #5