# Is There a Best Quality Metric for Graph Clusters?

Hélio Almeida [1], Dorgival Guedes [1], Wagner Meira Jr. [1], Mohammed J. Zaki [2]

[1]Universidade Federal de Minas Gerais

[2]Rensselaer Polytechnic Institute

September 5, 2011

inweb

# Graph Clustering

- Is the process of finding “communities” of similar vertices in a graph.
- Manually evaluating the quality of a given clustering is essential, but is hard, expensive and boring. Especially fo larger graphs.
- Quality metrics try to represent the most important cluster characteristics and can be used to evaluate its the fitness.

# The Problem

- Most papers just assume that a given chosen quality metric is good enough and run with it.
- There is no consensus on what is the best quality metric for graph clusters. Or even if it is possible to have a single best one.
- The lack of graphs (especially large ones) with known expected clusterings make it harder to evaluate the validity of clustering quality metrics in more complex/interesting cases.

# Our Contribution

- We wanted to verify if there in one quality metric that's markedly better than the others. If not, why?.
- We've chosen 5 popular structural quality metrics.
- Studied their structural characteristics. (Do they really represent good clusters?)
- Observed how they behave when applied to graphs with different sizes and origins. (Do they always behave as we expect?).
- Compared those metrics. (Do they agree on what is a good cluster? Is there a better clustering quality metric?)

# Quality Metrics Overview

- Clustering quality metrics aim to score a cluster (or whole clusterings) in terms of chosen characteristics that are believed to indicate well-formed clusters.
- Structurally speaking, a good cluster should have its vertices connected densely among themselves and sparsely with the rest of the graph.
- In this work, we've chosen 5 popular topological quality metrics:
  - Modularity.
  - Silhouette
  - Conductance
  - Coverage
  - Performance

# Modularity

- Measures the internal density and external sparsity of a given clustering.
- $Q$ is the fraction of all edges that lie within communities minus the expected value of the same quantity in a similarly built, albeit random, graph.

$$Q = Tr(e) - ||e^2||$$

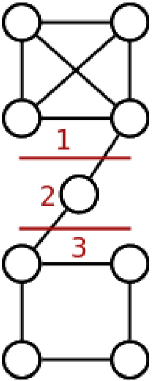
# Modularity

$\mathbf{e} =$

| | C1 | C2 | C3 |
|---|---|---|---|
| C1 | 0.5 | 0.08 | 0 |
| C2 | 0.08 | 0 | 0.08 |
| C3 | 0 | 0.08 | 0.33 |

$\mathbf{Q = 0.2999}$

- Singleton cluster (2): Is it that bad?



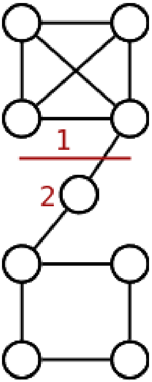
# Modularity

$\mathbf{e} =$

| | C1 | C2 |
|---|---|---|
| C1 | 0.5 | 0.08 |
| C2 | 0.08 | 0.4166 |

**Q = 0.3337**

- Is the new cluster 2 better than the old cluster 3?
- Is this clustering really better the the previous one? It only has less inter-cluster edges.



# Silhouette Index

- Uses vertex distances to measure cohesion and separation of clusters.
- A good cluster should have small average distance between its elements and greater average distance between them and other clusters.

# Silhouette Index

$$S_v = \frac{b_v - a_v}{max(a_v, b_v)}$$

- $a_v$: average distance between vertex $v$ and all other vertices in its own cluster.
- $b_v$: average distance between vertex $v$ and all vertices in the nearest cluster.
- Expensive (needs all-pairs shortest path calculation).
- Singleton clusters erroneously have high silhouette scores because $a_v = 0$.

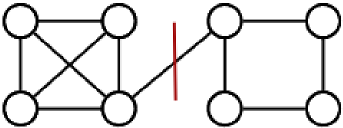
# Conductance

- The conductance of a graph cut measures its cost.
- If a clustering has low conductance value, it means that the clusters it defines are well separated. This concept is also called **intercluster (external) conductance**.
- If the graph induced by a cluster has high conductance, then it is too cohesive to be easily cut. This concept is also called **intracluster (internal) conductance**.
- Even though using both conductances would give better results, most authors ignore internal density because of its higher cost.

# Conductance

- External conductance is given by:

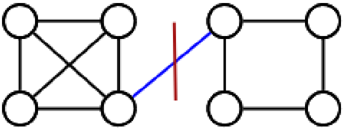
$$\phi(C_i) = \frac{\sum_{u \in C_i} \sum_{v \notin C_i} w(\{u, v\})}{min(a(C_i), a(\bar{C}_i))}$$



# Conductance

- External conductance is given by:

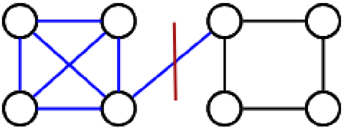
$$\phi(C_i) = \frac{\sum_{\mathbf{u}\in\mathbf{C_i}} \sum_{\mathbf{v}\notin\mathbf{C_i}} \mathbf{w}(\{\mathbf{u},\mathbf{v}\})}{min(a(C_i), a(\bar{C}_i))}$$



# Conductance

- External conductance is given by:

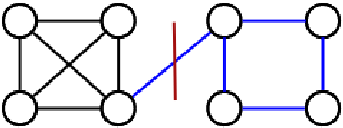
$$\phi(C_i) = \frac{\sum_{u \in C_i} \sum_{v \notin C_i} w(\{u, v\})}{min(\mathbf{a(C_i)}, a(\bar{C}_i))}$$



# Conductance

- External conductance is given by:

$$\phi(C_i) = \frac{\sum_{u \in C_i} \sum_{v \notin C_i} w(\{u, v\})}{min(a(C_i), \mathbf{a(\bar{C}_i)})}$$

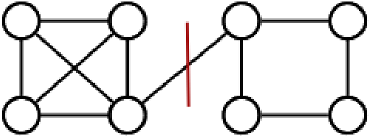


# Coverage

- It's the fraction of intracluster edges existent in the graph.
- High values of coverage mean that there are more edges inside the clusters than linking them, which is considered as a good clustering

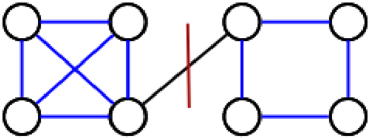
# Coverage

$$coverage(C) = \frac{w(C)}{w(G)}$$



# Coverage

$$coverage(C) = \frac{\mathbf{w(C)}}{w(G)}$$

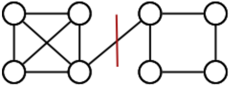


- Mainly uses inter-cluster sparsity to measure quality.
- Will be biased towards lower numbers of clusters.

# Performance

- Performance counts the number of edges linking vertices of a cluster among themselves, together with the number of edges that do **not** exist between them and the rest of the graph.
- High values mean that the cluster is both internally dense and externally sparse.

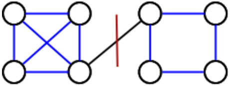
$$perf(C) = \frac{f(C) + g(C)}{\frac{1}{2}n(n-1)}$$



# Performance

- Performance counts the number of edges linking vertices of a cluster among themselves, together with the number of edges that do **not** exist between them and the rest of the graph.
- High values mean that the cluster is both internally dense and externally sparse.

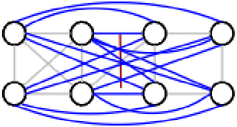
$$perf(C) = \frac{\mathbf{f(C)} + g(C)}{\frac{1}{2}n(n-1)}$$



# Performance

- Performance counts the number of edges linking vertices of a cluster among themselves, together with the number of edges that do **not** exist between them and the rest of the graph.
- High values mean that the cluster is both internally dense and externally sparse.

$$perf(C) = \frac{f(C) + \mathbf{g(C)}}{\frac{1}{2}n(n-1)}$$



# Performance

- Complex networks (especially social ones) tend to be sparse.
- In sparse graphs, the ratio of “nonexistent” edges will be way higher than the number of edges in the graph.
- Because of this, performance may lose its discerning power when applied to complex networks.

# Experiments Overview

- We wanted to compare the results of those quality metrics for different clusterings of real world graphs.
- To obtain different clusterings, we used 4 different clustering algorithms.
- We calculated the topological quality metrics discussed for each of those obtained clusterings.

# Clustering Algorithms Used

- Markovian (MCL)
- Bisecting K-means (CLUTO)
- Spectral (SCPS)
- Normalized Cut (GRACLUS)

# Datasets Used

| Network | # Vertices | # Edges |
|---|---|---|
| Karate Club | 34 | 78 |
| College Footbal | 115 | 616 |
| Astrophysics Collab. | 18772 | 396160 |
| H. E. Physics Collab. | 12008 | 237010 |
| H. E. Physics Citation | 34546 | 421578 |
| Gnutella Snap. (08/04/02) | 10876 | 39994 |
| Gnutella Snap. (08/30/02) | 36682 | 88328 |

# Discussion

- For the smaller graphs, communities found were very similar to the real ones.
- Metric values obtained are fairly good.
- Since it's a very small and popular dataset, this result is more than expected

# Astrophysics Collaboration Results

| Algorithm | # Clusters | SI | Mod. | Cover. | Perf. | Cond. |
|---|---|---|---|---|---|---|
| **MCL** | 1036 | -0.22 | 0.35 | 0.42 | 0.99 | 0.55 |
| | 2231 | -0.23 | 0.28 | 0.31 | 0.99 | 0.70 |
| | 4093 | 0.06 | 0.19 | 0.27 | 0.99 | 0.82 |
| **B. k-means** | 1037 | -0.73 | 0.25 | 0.28 | 0.99 | 0.70 |
| | 2232 | -0.48 | 0.21 | 0.24 | 0.99 | 0.70 |
| | 4094 | -0.21 | 0.17 | 0.19 | 0.99 | 0.76 |
| **Spectral** | 1034 | -0.15 | 0.34 | 0.38 | 0.99 | 0.53 |
| | 2131 | -0.26 | 0.25 | 0.28 | 0.99 | 0.66 |
| | 3335 | 0.04 | 0.19 | 0.21 | 0.99 | 0.78 |
| **Norm. Cut** | 1037 | -0.69 | 0.23 | 0.25 | 0.99 | 0.66 |
| | 2232 | -0.51 | 0.17 | 0.19 | 0.99 | 0.73 |
| | 4094 | -0.31 | 0.13 | 0.15 | 0.99 | 0.81 |

# Astrophysics Collaboration Results

| Algorithm | # Clusters | SI | Mod. | Cover. | Perf. | Cond. |
|---|---|---|---|---|---|---|
| **MCL** | 1036 | -0.22 | **0.35** | **0.42** | 0.99 | **0.55** |
| | 2231 | -0.23 | 0.28 | 0.31 | 0.99 | 0.70 |
| | 4093 | 0.06 | 0.19 | 0.27 | 0.99 | 0.82 |
| **B. k-means** | 1037 | -0.73 | **0.25** | **0.28** | 0.99 | **0.70** |
| | 2232 | -0.48 | 0.21 | 0.24 | 0.99 | 0.70 |
| | 4094 | -0.21 | 0.17 | 0.19 | 0.99 | 0.76 |
| **Spectral** | 1034 | -0.15 | **0.34** | **0.38** | 0.99 | **0.53** |
| | 2131 | -0.26 | 0.25 | 0.28 | 0.99 | 0.66 |
| | 3335 | 0.04 | 0.19 | 0.21 | 0.99 | 0.78 |
| **Norm. Cut** | 1037 | -0.69 | **0.23** | **0.25** | 0.99 | **0.66** |
| | 2232 | -0.51 | 0.17 | 0.19 | 0.99 | 0.73 |
| | 4094 | -0.31 | 0.13 | 0.15 | 0.99 | 0.81 |

# Astrophysics Collaboration Results

| Algorithm | # Clusters | SI | Mod. | Cover. | Perf. | Cond. |
|---|---|---|---|---|---|---|
| MCL | 1036 | -0.22 | 0.35 | 0.42 | 0.99 | 0.55 |
| | 2231 | -0.23 | 0.28 | 0.31 | 0.99 | 0.70 |
| | 4093 | **0.06** | 0.19 | 0.27 | 0.99 | 0.82 |
| B. k-means | 1037 | -0.73 | 0.25 | 0.28 | 0.99 | 0.70 |
| | 2232 | -0.48 | 0.21 | 0.24 | 0.99 | 0.70 |
| | 4094 | **-0.21** | 0.17 | 0.19 | 0.99 | 0.76 |
| Spectral | 1034 | -0.15 | 0.34 | 0.38 | 0.99 | 0.53 |
| | 2131 | -0.26 | 0.25 | 0.28 | 0.99 | 0.66 |
| | 3335 | **0.04** | 0.19 | 0.21 | 0.99 | 0.78 |
| Norm. Cut | 1037 | -0.69 | 0.23 | 0.25 | 0.99 | 0.66 |
| | 2232 | -0.51 | 0.17 | 0.19 | 0.99 | 0.73 |
| | 4094 | **-0.31** | 0.13 | 0.15 | 0.99 | 0.81 |

# Astrophysics Collaboration Results

| Algorithm | # Clusters | SI | Mod. | Cover. | Perf. | Cond. |
|---|---|---|---|---|---|---|
| MCL | 1036 | -0.22 | 0.35 | 0.42 | **0.99** | 0.55 |
| | 2231 | -0.23 | 0.28 | 0.31 | **0.99** | 0.70 |
| | 4093 | 0.06 | 0.19 | 0.27 | **0.99** | 0.82 |
| B. k-means | 1037 | -0.73 | 0.25 | 0.28 | **0.99** | 0.70 |
| | 2232 | -0.48 | 0.21 | 0.24 | **0.99** | 0.70 |
| | 4094 | -0.21 | 0.17 | 0.19 | **0.99** | 0.76 |
| Spectral | 1034 | -0.15 | 0.34 | 0.38 | **0.99** | 0.53 |
| | 2131 | -0.26 | 0.25 | 0.28 | **0.99** | 0.66 |
| | 3335 | 0.04 | 0.19 | 0.21 | **0.99** | 0.78 |
| Norm. Cut | 1037 | -0.69 | 0.23 | 0.25 | **0.99** | 0.66 |
| | 2232 | -0.51 | 0.17 | 0.19 | **0.99** | 0.73 |
| | 4094 | -0.31 | 0.13 | 0.15 | **0.99** | 0.81 |

# Gnutella Snapshot (08/04/02) Results

| Algorithm | # Clusters | SI | Mod. | Cover. | Perf. | Cond. |
|---|---|---|---|---|---|---|
| **MCL** | 2189 | -0.81 | 0.0004 | 0.001 | 0.99 | 0.99 |
| | 4724 | -0.037 | 0.0003 | 0.0007 | 0.99 | 0.99 |
| | 6089 | 0.1 | 0.00003 | 0.0003 | 0.99 | 1.00 |
| **B. k-means** | 2189 | -0.88 | 0.0004 | 0.001 | 0.99 | 0.99 |
| | 4724 | -0.52 | 0.00007 | 0.0004 | 0.99 | 0.99 |
| | 6089 | -0.18 | -0.00006 | 0.0002 | 0.99 | 1.00 |
| **Spectral** | 2158 | -0.90 | 0.0004 | 0.001 | 0.99 | 0.99 |
| | 4079 | -0.94 | 0.0001 | 0.0005 | 0.99 | 0.99 |
| | 6089 | -0.30 | -0.00007 | 0.0002 | 0.99 | 1.00 |
| **Norm. Cut** | 2189 | -0.90 | 0.0003 | 0.001 | 0.99 | 0.99 |
| | 4616 | -0.2 | 0.00025 | 0.0006 | 0.99 | 0.99 |
| | 5690 | 0.1 | 0.0002 | 0.0005 | 0.99 | 0.99 |

# Gnutella Snapshot (08/04/02) Results

| Algorithm | # Clusters | SI | Mod. | Cover. | Perf. | Cond. |
|---|---|---|---|---|---|---|
| **MCL** | 2189 | -0.81 | 0.0004 | 0.001 | **0.99** | 0.99 |
| | 4724 | -0.037 | 0.0003 | 0.0007 | **0.99** | 0.99 |
| | 6089 | 0.1 | 0.00003 | 0.0003 | **0.99** | 1.00 |
| **B. k-means** | 2189 | -0.88 | 0.0004 | 0.001 | **0.99** | 0.99 |
| | 4724 | -0.52 | 0.00007 | 0.0004 | **0.99** | 0.99 |
| | 6089 | -0.18 | -0.00006 | 0.0002 | **0.99** | 1.00 |
| **Spectral** | 2158 | -0.90 | 0.0004 | 0.001 | **0.99** | 0.99 |
| | 4079 | -0.94 | 0.0001 | 0.0005 | **0.99** | 0.99 |
| | 6089 | -0.30 | -0.00007 | 0.0002 | **0.99** | 1.00 |
| **Norm. Cut** | 2189 | -0.90 | 0.0003 | 0.001 | **0.99** | 0.99 |
| | 4616 | -0.2 | 0.00025 | 0.0006 | **0.99** | 0.99 |
| | 5690 | 0.1 | 0.0002 | 0.0005 | **0.99** | 0.99 |

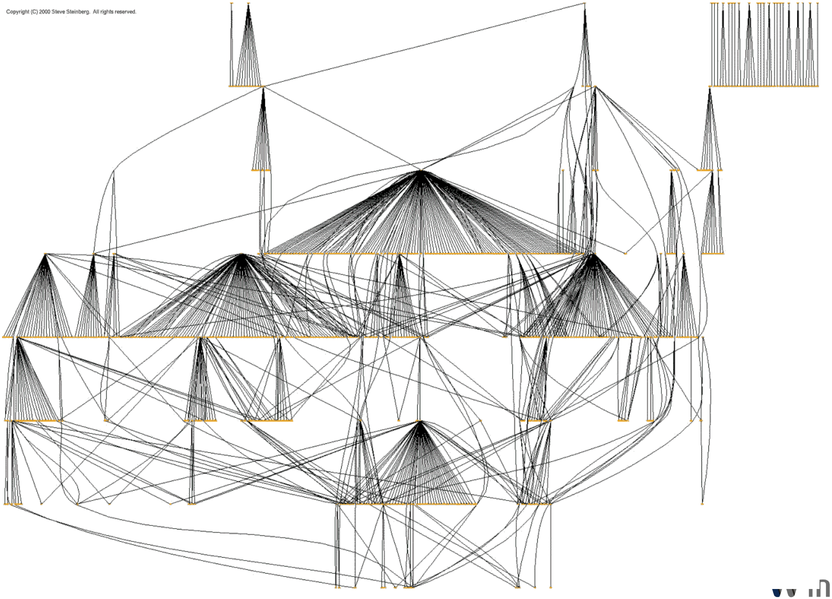
# Gnutella Snapshot (08/04/02) Results

| Algorithm | # Clusters | SI | Mod. | Cover. | Perf. | Cond. |
|---|---|---|---|---|---|---|
| **MCL** | 2189 | -0.81 | 0.0004 | 0.001 | 0.99 | 0.99 |
| | 4724 | -0.037 | 0.0003 | 0.0007 | 0.99 | 0.99 |
| | 6089 | **0.1** | 0.00003 | 0.0003 | 0.99 | 1.00 |
| **B. k-means** | 2189 | -0.88 | 0.0004 | 0.001 | 0.99 | 0.99 |
| | 4724 | -0.52 | 0.00007 | 0.0004 | 0.99 | 0.99 |
| | 6089 | **-0.18** | -0.00006 | 0.0002 | 0.99 | 1.00 |
| **Spectral** | 2158 | -0.90 | 0.0004 | 0.001 | 0.99 | 0.99 |
| | 4079 | -0.94 | 0.0001 | 0.0005 | 0.99 | 0.99 |
| | 6089 | **-0.30** | -0.00007 | 0.0002 | 0.99 | 1.00 |
| **Norm. Cut** | 2189 | -0.90 | 0.0003 | 0.001 | 0.99 | 0.99 |
| | 4616 | -0.2 | 0.00025 | 0.0006 | 0.99 | 0.99 |
| | 5690 | **0.1** | 0.0002 | 0.0005 | 0.99 | 0.99 |

# Gnutella Snapshot (08/04/02) Results

| Algorithm | # Clusters | SI | Mod. | Cover. | Perf. | Cond. |
|---|---|---|---|---|---|---|
| **MCL** | 2189 | -0.81 | **0.0004** | **0.001** | 0.99 | **0.99** |
| | 4724 | -0.037 | 0.0003 | 0.0007 | 0.99 | 0.99 |
| | 6089 | 0.1 | 0.00003 | 0.0003 | 0.99 | 1.00 |
| **B. k-means** | 2189 | -0.88 | **0.0004** | **0.001** | 0.99 | **0.99** |
| | 4724 | -0.52 | 0.00007 | 0.0004 | 0.99 | 0.99 |
| | 6089 | -0.18 | -0.00006 | 0.0002 | 0.99 | 1.00 |
| **Spectral** | 2158 | -0.90 | **0.0004** | **0.001** | 0.99 | **0.99** |
| | 4079 | -0.94 | 0.0001 | 0.0005 | 0.99 | 0.99 |
| | 6089 | -0.30 | -0.00007 | 0.0002 | 0.99 | 1.00 |
| **Norm. Cut** | 2189 | -0.90 | **0.0003** | **0.001** | 0.99 | **0.99** |
| | 4616 | -0.2 | 0.00025 | 0.0006 | 0.99 | 0.99 |
| | 5690 | 0.1 | 0.0002 | 0.0005 | 0.99 | 0.99 |

# Example of Gnutella Network Topology



# Discussion

- The network structure has a very low probability of generating clusters as expected from the quality metrics.
- Probability of 3-clique occurrence is only 0.5%, while it is 31.8% for the Astrophysics collaboration network, for example.
- Also, by design, Gnutella networks are *very* sparse.
- Only 6.76% of all possible edges in fact exist in this Gnutella snapshot (opposed to 32.88% for the H. E. Physics citation network, for example).

# Conclusions

- The quality metrics studied do not share a common view on what is a good clustering.
- They present strong biases that do not necessarily indicate good clusters.
- Graphs of different origins might have different characteristics and, therefore, have different cluster structure signatures.
- From all that, we concluded that none of those quality metrics represents the characteristics of a well-formed cluster with a good degree of precision.

# Future Work

- New, more adequate graph clustering quality metrics are needed.
- Study large networks to identify how its characteristics influence cluster structures.
- Also, study how other information dimensions (such as edge weights and asymmetry or vertex labels) affect cluster structures.

# The end.

Questions?