# A Statistical Performance Analysis of Graph Clustering Algorithms

WAW 2018, 15th Workshop on Algorithms and Models for the Web Graph, Moscow Institute of Physics and Technology, Dolgoprudny, Russia, May 18, 2018

P. Miasnikof [1] A. Y. Shestopaloff [2] A. J. Bonner [3] Y. Lawryshyn [1]

University of Toronto, Dept. of Chemical Engineering and Applied Chemistry, Toronto, Canada

The Alan Turing Institute, London, United Kingdom

University of Toronto, Dept. of Computer Science, Toronto, Canada

# Acknowledgements

- PM thanks Derek Corneil, Liudmila Ostroumova Prokhorenkova, Mark Newman, Cris Moore, Aaron Clauset, Anne Morvan and Leonidas Pitsoulis for their helpful comments on this work. AND the organizers for their invitation!
- PM was supported by Mitacs-Accelerate PhD award IT05806.

# Outline

# Disclaimers

While there are subtle differences between the two concepts, we use “graph clustering” & “network community detection” interchangeably.

# Disclaimers

While there are subtle differences between the two concepts, we use "graph clustering" & "network community detection" interchangeably.

Here, we do not perform clustering, our goal is limited to measuring the quality of clusterings returned by any clustering technique. Our measures are algorithm agnostic.

# Motivations I

- Graph clustering (and network community detection) is a topic that has received much attention
- There are quite a few clustering algorithms in the literature, but assessing their performance objectively remains an open problem

# Motivations II

- "*A cluster is defined as whatever is returned by a clustering algorithm*" - (LOP's ironic statement at WAW 2017)
- "*(...) running a clustering algorithm over a set of randomly generated data points will always produce clusters which, however, have little meaning.*"
  [Reichardt and Bornholdt, 2006]

# Motivations III

- Reliance on "ground truth" data sets does indeed provide objective reproducible performance measurements
- Unfortunately, it does not provide guarantees the algorithm will perform similarly well on another data set

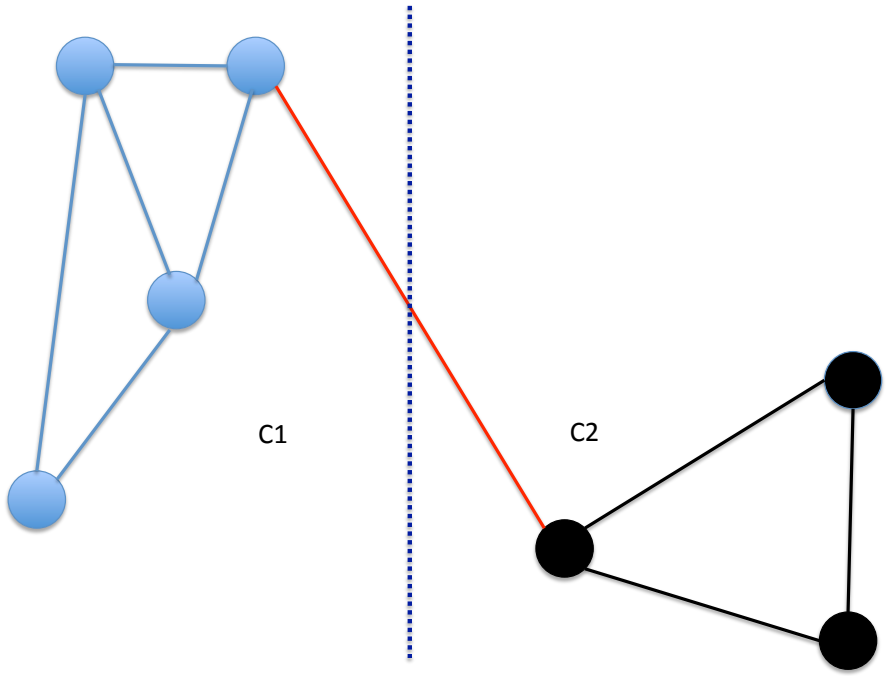
# Objectives for This Presentation

- **Stimulate discussion & gather your feedback!**
- Define our own understanding of a good clustering
- Introduce new performance measures
- Describe empirical comparisons to other measures from the literature

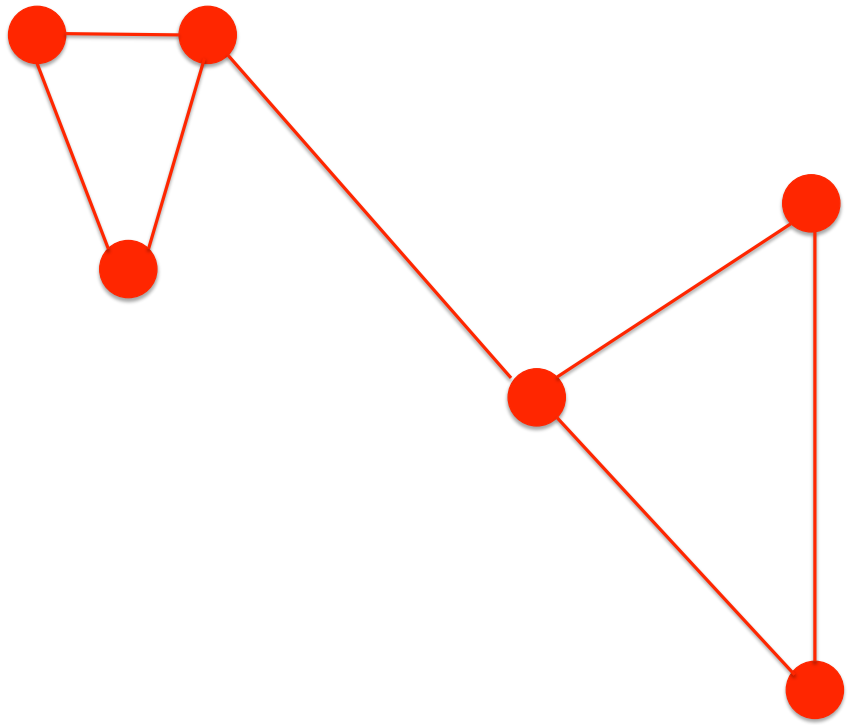
# Characteristics of a (Graph) Cluster

- A cluster (or community) is a subset of vertices that exhibit [Yang and Leskovec, 2012] (we quote these authors, but their definition is very common throughout the literature)
  - High-level of interconnection between vertices within itself
  - Low-level of connection to vertices in the rest of the graph
  - Not necessarily cliques
- A graph may or may not exhibit a clustered structure

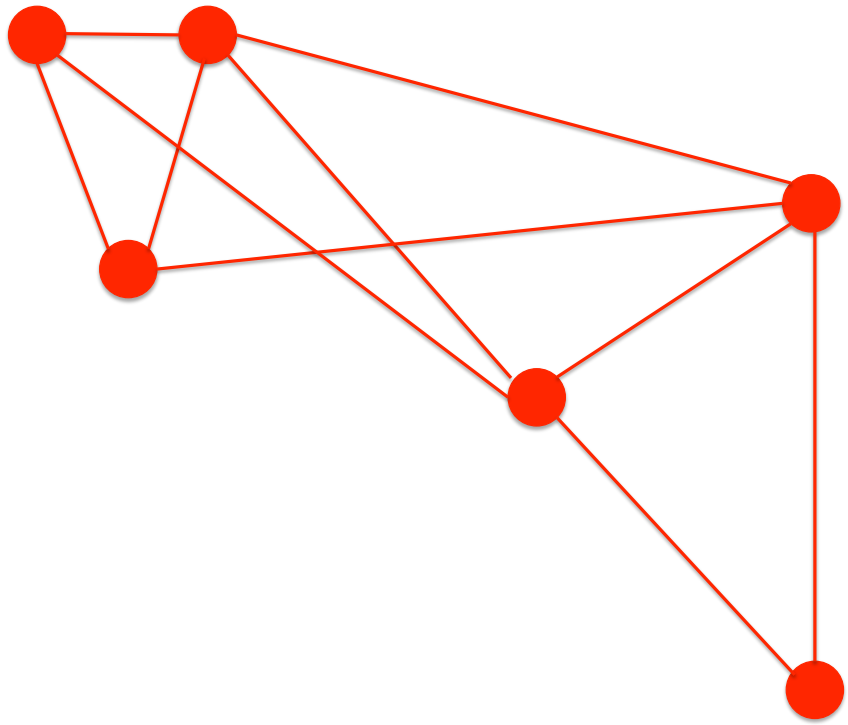
# An Example of a Good Clustering



# An Example of a Bad Clustering



# An Example of a Graph Without Obvious Clusters



# What Do Our Measures Strive to Achieve?

- Our statistical benchmarks are an attempt to formally measure the strength of the clusters, across the graph
- Our goal is to formally test the null hypothesis that the algorithm's clustering is the result of a random assignment
- Our measures can be used to determine if the cluster labeling returned by a specific algorithm is of good quality and to compare the quality of clusterings identified by two (or more) different algorithms

# Defining Our Variables I

- The set of all clusters: $C = \{C_1, \ldots, C_\kappa\}$, with $|C| = \kappa$
- Total number of vertices in the graph: $N$
- Total number of vertices in cluster $i$: $|C_i| = n_i$
- $\sum_{i=1}^{\kappa} n_i = N$

# Defining Our Variables II

- The set of all edges on the graph: $E = \{e_1, \ldots, e_m\}$
- The sets of edges connecting two vertices in cluster $i$: $E_{i,i} \in \{E_{1,1}, E_{2,2}, \ldots, E_{\kappa,\kappa}\}$
- The sets of edges connecting two vertices in two different clusters $C_i$ and $C_j$: $E_{i,j} \in \{E_{1,2}, \ldots, E_{1,\kappa}, \ldots, E_{\kappa-1,\kappa}\}$, where $i, j \in [1, \kappa],\ i \neq j$

# Clustering Statistics I

The **MAIN** benchmark

- Graph's connections ratio ($\in [0, 1]$):

$$\bar{K} = \frac{|E|}{0.5 \times N(N-1)}$$

# Clustering Statistics II

- Mean intra-cluster connections ratio ($\in [0, 1]$):

$$\bar{K}_{\text{intra}} = \frac{1}{\kappa} \sum_{i=1}^{\kappa} \frac{|E_{i,i}|}{0.5 \times n_i(n_i - 1)}$$

# Clustering Statistics III

- Mean inter-cluster connections ratio ($\in [0,1]$):

$$\bar{K}_{\text{inter}} = \frac{1}{0.5 \times \kappa(\kappa-1)} \times \sum_{i=1}^{\kappa} \sum_{j=i+1}^{\kappa} \frac{|E_{i,j}|}{0.5 \times ((n_i+n_j)(n_i+n_j-1) - n_i(n_i-1) - n_j(n_j-1))}$$

# Clustering Statistics IV

**WHY** is $\bar{K}$ the **MAIN** benchmark?

- Imagine $\bar{K}_{\text{intra}} = 0.86$
  - BUT $\bar{K} = 0.89$ and $\bar{K}_{\text{inter}} = 0.89$
- Now imagine $\bar{K}_{\text{intra}} = 0.70$
  - BUT $\bar{K} = 0.61$ and $\bar{K}_{\text{inter}} = 0.60$
- Or the exttreme case of a complete graph
  - $\bar{K}_{\text{intra}} = \bar{K} = \bar{K}_{\text{inter}} = 1$
- If the clustering quality is good, we expect the inequalities $\bar{K}_{\text{intra}} > \bar{K} > \bar{K}_{\text{inter}}$ to hold at a reasonable significance level

# Statistical interpretation I

- The main strength of our Kappas comes from their statistical definition
- Our measures are defined as statistical measurements with associated standard errors, not deterministic quantities
- Our statistical (i.e., non deterministic) definition also allows for uncertainty in the connectivity, another open problem [Holder et al., 2016]

# Statistical interpretation II

Our measures can be interpreted as empirical estimates of the probability two vertices are connected

- $\bar{K}$: probability any two vertices are connected by an edge (i.e., $P(e_{i,j})$)
- $\bar{K}_{\text{intra}}$: conditional probability any two vertices within the same cluster are connected by an edge (i.e., $P(e_{i,j}|c_i = c_j)$)
- $\bar{K}_{\text{inter}}$: conditional probability any two vertices in different clusters are connected by an edge (i.e., $P(e_{i,j}|c_i \neq c_j)$)

# Null Hypotheses

- $H_o^{(a)}$: The clustering is the result of a random assignment, therefore the mean intra-cluster connections ratio and the graph's connections ratio are statistically indistinguishable, or $\bar{K}_{\text{intra}} \leq \bar{K}$

# Null Hypotheses

- $H_o^{(a)}$: The clustering is the result of a random assignment, therefore the mean intra-cluster connections ratio and the graph's connections ratio are statistically indistinguishable, or $\bar{K}_{\text{intra}} \leq \bar{K}$
- $H_o^{(b)}$: The clustering is the result of a random assignment, therefore the mean inter-cluster connections ratio and the graph's connections ratio are statistically indistinguishable, or $\bar{K}_{\text{inter}} \geq \bar{K}$

# Hypothesis Tests

To test $H_o^{(a)}$, we use a $t$ distribution with $\kappa - 1$ degrees of freedom. The test statistic in this case is

$$t_a = \frac{\bar{K}_{\text{intra}} - \bar{K}}{se(\text{Mean intra-cluster connections ratio})}$$

# Hypothesis Tests

To test $H_o^{(a)}$, we use a $t$ distribution with $\kappa - 1$ degrees of freedom. The test statistic in this case is

$$t_a = \frac{\bar{K}_{\text{intra}} - \bar{K}}{se(\text{Mean intra-cluster connections ratio})}$$

To test $H_o^{(b)}$, we use a $t$ distribution with $0.5\kappa(\kappa - 1) - 1$ degrees of freedom. The test statistic in this case is

$$t_b = \frac{\bar{K}_{\text{inter}} - \bar{K}}{se(\text{Mean inter-cluster connections ratio})}$$

# Experimental Design I

- We generated a set of ***contrived simplistic*** examples to examine each performance measure's sensitivity to graph structure (Learn to walk walk, before you learn to ride a bike)
- We varied inter and intra cluster connectivity, while maintaining cluster labeling constant
- We then computed each metric on graphs where the cluster labelling reflected a good partition and where it did not
- Proceeding in this way allowed us to simulate output of good & bad clustering algorithms

# Experimental Design II

An example with two clusters: varying intra-cluster connectivity with no noise from inter-cluster connectivity.

$$\begin{bmatrix} 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \end{bmatrix} \rightarrow \dots \rightarrow \begin{bmatrix} 1 & \dots & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 1 & \dots & 1 \end{bmatrix}$$

# Experimental Design III

We then varied inter-cluster connectivity (still without noise):

$$\begin{bmatrix} 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \end{bmatrix} \rightarrow \dots \rightarrow \begin{bmatrix} 0 & \dots & 0 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 1 & \dots & 1 \\ 1 & \dots & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \dots & 1 & 0 & \dots & 0 \end{bmatrix}$$

They were then repeated with the addition of noise from intra/inter-cluster connectivity

# Results I

| **Pct Inter = 0, Pct Intra varies** | | | | | |
|---|---|---|---|---|---|
| **Pct Intra** | 0 | 25 | 50 | 75 | 100 |
| **N** | 10,048 | 9,440 | 9,666 | 10,493 | 10,039 |
| $\|\mathbf{C}\|$ | 200 | 200 | 200 | 200 | 200 |
| $\|\mathbf{E}\|$ | 0 | 76,942 | 160,147 | 269,341 | 336,942 |
| $\bar{K}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 |
| $\bar{K}_{intra}$ | 0.00 | 0.26 | 0.50 | 0.75 | 0.99 |
| **Std Err** $(\bar{K}_{intra})$ | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 |
| $\bar{K}_{inter}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **Std Err** $(\bar{K}_{inter})$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $\Phi$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **Q** | 0.00 | 0.99 | 0.99 | 0.99 | 0.99 |

# Results II

| **Pct Intra = 0, Pct Inter varies** | | | | | |
|---|---|---|---|---|---|
| **Pct Inter** | 0 | 25 | 50 | 75 | 100 |
| **N** | 10,530 | 10,089 | 9,354 | 10,028 | 10,829 |
| $\|\mathbf{C}\|$ | 200 | 200 | 200 | 200 | 200 |
| $\|\mathbf{E}\|$ | 0 | 3,058,924 | 10,753,463 | 27,815,367 | 58,250,108 |
| $\bar{K}$ | 0.00 | 0.06 | 0.25 | 0.55 | 0.99 |
| $\bar{K}_{intra}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **Std Err** $(\bar{K}_{intra})$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $\bar{K}_{inter}$ | 0.00 | 0.06 | 0.24 | 0.55 | 1.00 |
| **Std Err** $(\bar{K}_{inter})$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $\Phi$ | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| **Q** | 0.00 | -0.01 | -0.01 | -0.01 | -0.01 |

# A Closer Look at Modularity

Here, $|C| = 200, N = 16,400$ and $n_i = 82\ \forall i$

| Scenarios | A0 | | A25 | | A100 | |
|---|---|---|---|---|---|---|
| Components of Q | e_ii | a_i | e_ii | a_i | e_ii | a_i |
| cluster 1 | 0 | 0.005 | 0.00001 | 0.005 | 0.00002 | 0.005 |
| cluster 2 | 0 | 0.005 | 0.00001 | 0.005 | 0.00002 | 0.005 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| cluster K | 0 | 0.005 | 0.00001 | 0.005 | 0.00002 | 0.005 |

Added intra-cluster connectivity only has an infinitesimal effect on the value of Q... (Recall: $Q = \sum_i (e_{i,i} - a_i^2) \rightarrow Q = 200 \times (0 - 0.005^2) \approx 0$) By contrast, $\bar{K}_{\text{intra}}$ picked up this added connectivity (please see Table 3 in paper)

# Summary

- We offered a statistical definition of clustering quality
- We formulated stat hypotheses & formal tests to assess them
- Our measures are more reflective of the graph's underlying clustered structure

# Future Work

- Look into theoretical properties of quality functions (e.g., [Kehagias and Pitsoulis, 2018])
- Extend to overlapping clusters
- Apply to more complex graph structures
- Explore alternatives to connectivity as performance metrics (e.g., distance)

**THANKS FOR YOUR TIME!**

**QUESTIONS? COMMENTS?**

**A LOT MORE IN THE PAPER! PLEASE READ IT! :-)**

**contact: p.miasnikof@mail.utoronto.ca**


Holder, L. B., Caceres, R., Gleich, D. F., Riedy, J., Khan, M., Chawla, N. V., Kumar, R., Wu, Y., Klymko, C., Eliassi-Rad, T., and Prakash, A. (2016).
Current and future challenges in mining large networks: Report on the second sdm workshop on mining networks and graphs.
*SIGKDD Explor. Newsl.*, 18(1):39–45.


Kehagias, A. and Pitsoulis, L. (2018).
Graph Clustering Quality Functions.
*Unpublished*.


Reichardt, J. and Bornholdt, S. (2006).
When are networks truly modular?
*Physica D: Nonlinear Phenomena*, 224(1):20 – 26.
Dynamics on Complex Networks and Applications.


Yang, J. and Leskovec, J. (2012).

Defining and Evaluating Network Communities based on Ground-truth.
*CoRR*, abs/1205.6233.