Distances on a Graph

Pierre Miasnikof[1] ⋆, Alexander Y. Shestopaloff[2,3], Leonidas Pitsoulis[4], Alexander Ponomarenko[5], and Yuri Lawryshyn[1]

[1] University of Toronto, Toronto, ON, Canada,
[2] Queen Mary University of London, London, United Kingdom,
[3] The Alan Turing Institute, London, United Kingdom,
[4] Aristotle University of Thessaloniki, Thessaloniki, Greece
[5] National Research University Higher School of Economics, Laboratory of Algorithms and Technologies for Networks Analysis, Nizhny Novgorod, Russian Federation

Abstract. In this article, our ultimate goal is to transform a graph's adjacency matrix into a distance matrix. Because cluster density is not observable prior to the actual clustering, our goal is to find a distance whose pairwise minimization will lead to densely connected clusters. Our thesis is centered on the widely accepted notion that strong clusters are sets of vertices with high induced subgraph density. We posit that vertices sharing more connections are closer to each other than vertices sharing fewer connections. This definition of distance differs from the usual shortest-path distance. At the cluster level, our thesis translates into low mean intra-cluster distances, which reflect high densities. We compare three distance measures from the literature. Our benchmark is the accuracy of each measure's reflection of intra-cluster density, when aggregated (averaged) at the cluster level. We conduct our tests on synthetic graphs, where clusters and intra-cluster density are known in advance. In this article, we restrict our attention to unweighted graphs with no self-loops or multiple edges. We examine the relationship between mean intra-cluster distances and intra-cluster densities. Our numerical experiments show that Jaccard and Otsuka-Ochiai offer very accurate measures of density, when averaged over vertex pairs within clusters.

1 Introduction

When clustering graphs, we seek to group nodes into clusters of nodes that are similar to each other. We posit that similarity is reflected in the number of shared connections. Our node-to-node distances are based on this shared connectivity. Although a formal definition of vertex clusters (communities) remains a topic of debate, virtually all authors agree a cluster is a subset of vertices that exhibit a high level of interconnection between themselves and a low level of connection to vertices in the rest of the graph [21, 7, 19, 20] (we quote these authors, but their definition is very common across the literature). Consequently, clusters, subsets of strongly inter-connected vertices, also form dense induced subgraphs.

⋆ Corresponding author: p.miasnikof@mail.utoronto.ca

Unfortunately, cluster density is not observable prior to the actual clustering. For this reason, we want a quantity that guides the aggregation of vertices into clusters, so that they form pockets of vertices separated by smaller than average distances, pockets of highly inter-connected vertices. To this end, we compare the accuracy of various node-to-node distance measures in reflecting intra-cluster density.

2 Distance, Intra-Cluster Density and Graph Clustering (Network Community Detection)

As mentioned previously, clusters are defined as subsets of vertices that are considered somehow similar. This similarity is captured by the number of shared connections and translated into distance. In our model, vertices sharing a greater number of connections are separated by smaller distances than vertices with which they share fewer connections. It is important to note here that, in our definition, distance measures similarity, not geodesic (shortest path) distance. For example, two vertices with high degrees that share an edge but no other connection have a geodesic distance of one, but are dissimilar on the basis of their connectivity. At the cluster level, smaller within-cluster distances reflect subsets of more densely connected vertices.

In this article, our ultimate goal is to transform a graph's adjacency matrix into a $|V| \times |V|$ distance matrix $D = [d_{ij}]$, where the distance between each pair of vertices is given by the element $d_{ij}(\geq 0)$. This transformation allows us to cluster using distance minimization techniques from the literature. The quadratic formulation of Fan and Pardalos [6, 5] and the K-medoids technique [2] are examples of such graph clustering techniques. These formulations can then be further modified into a QUBO formulation [10]. This reformulation can then be solved using newly available purpose-built hardware which allows us to circumvent the NP-hardness of the clustering problem [9, 21, 7, 14, 1].

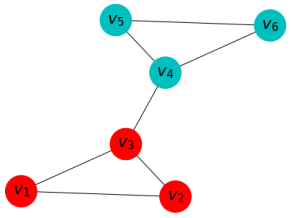

Fig. 1. Graph with Two Clusters

To illustrate our definition of distance, we examine the graph shown in Figure 1. The graph in that figure is arguably composed of two clusters (triangles), the red cluster containing vertices v_1, v_2, v_3 and the cyan cluster with vertices v_4, v_5, v_6. We observe that each cluster forms a dense induced subgraph (clique).

We also note that the geodesic distance separating vertices v_1 and v_3 is equal to the geodesic distance separating v_3 and v_4. Nevertheless, in the context of clustering, we argue that v_3 is closer, more similar, to v_1 than to v_4.

3 Distance Measurements Under Study

We compare three different distance measurements from the literature and examine how faithfully they reflect connectivity patterns. We argue that mean node-to-node distance within a cluster should offer an accurate reflection of intra-cluster density, but move in an opposite direction. Densely connected clusters should display low mean node-node distances.

Intra-cluster density is defined as

$$K_{\text{intra}}^{(k)} = \frac{|E_{kk}|}{0.5 \times n_k \times (n_k - 1)} \,.$$

In this definition, $|E_{kk}|$ is the cardinality of the set of edges connecting two vertices within the same cluster 'k' and $n_k = |V_k|$ is the number of vertices in that same cluster. This ratio also represents the empirical estimate of the probability two nodes within a cluster are connected by an edge.

We then examine the relationship between mean Jaccard [12], Otsuka-Ochiai [17] and Burt's distances [3, 7], on one hand, and intra-cluster density within each cluster, on the other. Because these distances are pairwise measures, we compare their mean value for a given cluster to the cluster's internal density.

3.1 Embedding, Commute and Amplified Commute Distances

We begin by calling the reader's attention to the fact this article is not about graph embedding. Here, we are not interested in a vector representation of nodes. We are only interested in the distance separating them.

We also call the reader's attention to the fact the distance measures under consideration can all be obtained using simple arithmetic. It is precisely for this reason that we did not consider the popular "commute distance" and its corrections, like "amplified commute distance" [22, 23, 18], in this work. While these distances are known to capture cluster structure, they require matrix inversion and are very costly to compute [4]. Although some authors have found efficient approximations that circumvent the need for matrix inversion (e.g., [22]), the distances under consideration in this article are exact quantities. Exactness of the distances is a desirable feature, given our ultimate goal to use them to estimate intra-cluster density. Additionally, unlike some of the approximations in the literature, our distances have simple and intuitive interpretations.

3.2 Jaccard Distance

The Jaccard distance separating two vertices 'i' and 'j' is defined as

$$\zeta_{ij} = 1 - \frac{|c_i \cap c_j|}{|c_i \cup c_j|} \in [0, 1] \,.$$

Here, $c_i\,(c_j)$ represents the set of all vertices with which vertex '$i\,(j)$' shares an edge.

At the cluster level, we compute the mean distance separating all pairs of vertices within the cluster, which we denote as $\mathcal{J}$. For an arbitrary cluster 'k' with n_k vertices, we have

$$\mathcal{J}_k = \frac{1}{0.5 \times n_k \times (n_k - 1)} \sum_{i,j=i+1} \zeta_{ij}\,.$$

3.3 Otsuka-Ochiai Distance

The Otsuka-Ochiai (OtOc) distance separating two vertices 'i' and 'j' is defined as

$$o_{ij} = 1 - \frac{|c_i \cap c_j|}{\sqrt{|c_i| \times |c_j|}} \in [0, 1]\,.$$

Here too, we obtain a cluster level measure of similarity by taking the mean over each pair of nodes within a cluster. We denote this mean as $\mathcal{O}$. Again, for an arbitrary cluster 'k' with n_k vertices, we have

$$\mathcal{O}_k = \frac{1}{0.5 \times n_k \times (n_k - 1)} \sum_{i,j=i+1} o_{ij}\,.$$

3.4 Burt's Distance

Burt's distance between two vertices 'i' and 'j', denoted as b_{ij}, is computed using the adjacency matrix (A) as

$$b_{ij} = \sqrt{\sum_{k \neq i,j} \left(A_{ik} - A_{jk}\right)^2}\,.$$

At the cluster level, we denote the mean Burt distance as $\mathcal{B}$. As with the other distances, for an arbitrary cluster 'k' with n_k vertices, it is computed as

$$\mathcal{B}_k = \frac{1}{0.5 \times n_k \times (n_k - 1)} \sum_{i,j=i+1} b_{ij}\,.$$

4 Numerical Comparisons

To compare the distance measures and assess the accuracy of each measure as a reflection of intra-cluster density, we generate synthetic graphs with known cluster membership, using the NetworkX library's [11] stochastic block model generator. In our experiments, we generate several graphs with varying graph and cluster sizes and inter and intra-cluster edge probabilities. To ensure ease of readability, we only include a subset of our most revealing results.

For each test graph in the experiments below, we compute our three vertex-to-vertex distances. We then compute mean distances between nodes in each cluster and intra-cluster density. To obtain a graph-wide assessment, we then take the mean of all cluster quantities over the entire graph. Because our clusters vary in size, we ensure the well-documented "resolution limit" degeneracy [8] does not affect our conclusions by taking simple unweighted means, regardless of cluster sizes.

4.1 Test Data: Synthetic Graphs with Known Clusters

We use the stochastic block model to generate two sets of six graphs, as described in Table 1. In the first set of experiments, we vary the probability of an intra-cluster edge, an edge with both ends inside the cluster. To generate noise, we vary the size (n_k) and number of clusters (K) and as a result the total number of nodes ($|V|$). For added noise, we also set inter-cluster edge probability to 0.15. Details are shown in Table 1.

Table 1. Synthetic Graphs and their Characteristics

First Set of Experiments					
Graph	Intra Pr	Inter Pr	K	n_k	$\|V\|$
G1	1	0.15	39	[38, 77]	3,641
G2	0.8	0.15	47	[38, 77]	4,703
G3	0.6	0.15	47	[38, 77]	4,326
G4	0.4	0.15	55	[38, 77]	5,386
G5	0.2	0.15	56	[38, 77]	5,557
G6	0	0.15	39	[38, 77]	3,705
Second Set of Experiments					
G7	1	0.15	60	[38, 77]	3,400
G8	0.8	0.15	60	[38, 77]	3,400
G9	0.6	0.15	60	[38, 77]	3,400
G10	0.4	0.15	60	[38, 77]	3,400
G11	0.2	0.15	60	[38, 77]	3,400
G12	0	0.15	60	[38, 77]	3,400

In the second set of experiments, in order to isolate the effect of intra-cluter edge probability, we keep the total number of clusters, nodes in each cluster and, consequently, total number of nodes fixed across all graphs. Although our cluster sizes vary within the graph, they are kept constant in each graph. Clusters $c_1, \ldots, c_K$ in graphs $G7, \ldots, G12$ all have $n_1, \ldots, n_K$ nodes. In this experiment, we only vary intra-cluster edge probability. Details are also shown in Table 1.

4.2 Empirical Results

As mentioned earlier, we have conducted several experiments with varying graph and cluster sizes and inter and intra-cluster edge probabilities. In the interest of brevity, we only present the most illustrative subset of our results.

In our first set of experiments, we begin by observing that our results confirm intra-cluster density is a very accurate estimator of intra-cluster edge probability, under all scenarios. This observation is consistent with prior work linking densities and clustering [15, 16]. We also note that both Jaccard and OtOc distances offer a good reflection of intra-cluster density and that their change under variations in intra-cluster edge probability are in reversed lock-step with intra-cluster density. Finally, we note Burt's distance offers a poor reflection of intra-cluster density. These results are shown in Table 2.

Table 2. First Set of Graph Experiments (G1-G6)

	P_intra					
	0	0.2	0.4	0.6	0.8	1
Jacc						
mean	0.919	0.918	0.912	0.898	0.879	0.841
stdev	0.000	0.000	0.002	0.005	0.012	0.020
+1 stdev	0.919	0.919	0.914	0.903	0.891	0.862
-1 stdev	0.919	0.918	0.910	0.893	0.867	0.821
OtOc						
mean	0.850	0.849	0.838	0.815	0.784	0.727
stdev	0.001	0.001	0.003	0.008	0.019	0.030
+1 stdev	0.851	0.849	0.842	0.823	0.803	0.757
-1 stdev	0.849	0.848	0.835	0.807	0.765	0.697
Burt						
mean	30.325	37.732	37.355	33.499	34.708	30.059
stdev	0.125	0.062	0.101	0.095	0.055	0.138
+1 stdev	30.450	37.794	37.455	33.594	34.763	30.197
-1 stdev	30.200	37.671	37.254	33.404	34.653	29.921
K_intra						
mean	0.000	0.199	0.400	0.601	0.800	1.000
stdev	0.000	0.005	0.007	0.008	0.006	0.000
+1 stdev	0.000	0.204	0.407	0.609	0.806	1.000
-1 stdev	0.000	0.195	0.393	0.593	0.794	1.000

In our second set of experiments, shown in Table 3, we observe the same relationship between distances and density. However, we also observe a factor of two reduction in the noise of both Jaccard and OtOc distances, while Burt's distance remains roughly at the same level of noise in both sets of experiments. A more detailed examination of this noise phenomenon is provided in the next section.

Table 3. Second Set of Graph Experiments (G7-G12)

	P_intra					
	0	0.2	0.4	0.6	0.8	1
Jacc						
mean	0.919	0.918	0.913	0.903	0.888	0.868
stdev	0.000	0.001	0.001	0.003	0.006	0.010
+1 stdev	0.919	0.919	0.914	0.906	0.894	0.878
-1 stdev	0.919	0.918	0.911	0.899	0.882	0.858
OtOc						
mean	0.850	0.849	0.840	0.823	0.798	0.767
stdev	0.001	0.001	0.002	0.005	0.010	0.015
+1 stdev	0.851	0.850	0.842	0.828	0.808	0.783
-1 stdev	0.849	0.848	0.837	0.817	0.788	0.752
Burt						
mean	29.190	29.496	29.650	29.641	29.485	29.205
stdev	0.068	0.073	0.071	0.076	0.061	0.103
+1 stdev	29.258	29.569	29.721	29.717	29.546	29.308
-1 stdev	29.122	29.422	29.579	29.566	29.423	29.102
K_intra						
mean	0.000	0.200	0.402	0.599	0.800	1.000
stdev	0.000	0.011	0.015	0.011	0.011	0.000
+1 stdev	0.000	0.211	0.417	0.610	0.811	1.000
-1 stdev	0.000	0.189	0.387	0.588	0.788	1.000

4.3 Noise, Sensitivity and Convergence

To better understand the sensitivity of each distance to variations in intra-cluster edge probability, we examine their asymptotic convergence. Using their definitions, we study their behavior as intra-cluster edge probability approaches 0 or 1, while keeping all else equal.

For each examination below, we define the following variables:

- P_i: probability of intra-cluster edge
- P_o: probability of inter-cluster edge
- N: total number of nodes on the graph
- n_k: number of nodes in an arbitrary cluster k
- c_i, c_j: the set of connections of two arbitrary vertices i, j in the same cluster
- A: the graph's adjacency matrix

Jaccard (and OtOc)

$$\begin{aligned}\zeta_{ij} &= 1 - \frac{|c_i \cap c_j|}{|c_i \cup c_j|} \\ &\approx 1 - \frac{P_i^2 \times (n_k - 2) + P_o^2 \times (N - n_k)}{P_i \times (n_k - 2) + P_o \times (N - n_k)}\end{aligned}$$

From this definition, we observe that

$$P_i \to 0 \Rightarrow \zeta_{ij} \to 1 - \frac{P_o^2 \times (N - n_k)}{P_o \times (N - n_k)}$$
$$P_i \to 1 \Rightarrow \zeta_{ij} \to 1 - \frac{(n_k - 2) + P_o^2 \times (N - n_k)}{(n_k - 2) + P_o \times (N - n_k)}\,.$$

The main observation here is that while the actual Jaccard distance depends on the number of nodes in each cluster and the total number of nodes on the graph, its variation remains in step with intra-cluster edge probability and intra-cluster density. It is this dependence on the number of nodes in each cluster and the total number of nodes on the graph that is the main source of additional variance observed in Table 2 and which is mitigated by keeping cluster sizes constant across graphs in the second set of experiments shown in Table 3. A similar argument can be made in the case of OtOc.

Burt's Distance

$$b_{ij} = \sqrt{\sum_{k \neq i,j} (A_{ik} - A_{jk})^2}$$
$$\approx \sqrt{2 \times P_i(1 - P_i) \times (n_k - 2) + 2 \times P_o(1 - P_o) \times (N - n_k)}$$

From this definition, we observe that

$$P_i \to 0 \Rightarrow b_{ij} \to \sqrt{2 \times P_o(1 - P_o) \times (N - n_k)}$$
$$P_i \to 1 \Rightarrow b_{ij} \to \sqrt{2 \times P_o(1 - P_o) \times (N - n_k)}\,.$$

On the other hand, the asymptotic behavior of Burt's distance explains why it is a poor reflection of intra-cluster density. We see that as P_i moves toward either extreme, Burt's distance moves toward the same quantity. It should also be noted that it is unbounded and grows with the number of nodes on the graph. In fact, as the total number of nodes increases in proportion to cluster size, the intra-cluster portion is minimized, since $(n_k - 2) \ll (N - n_k)$.

5 Our Chosen Distance

Both Jaccard and OtOc distances are very accurate reflections of intra-cluster density. Clustering by minimizing either will result in dense clusters. However, the Jaccard distance displays lower variance, in our experiments. Additionally, Jaccard similarity and its complement, the Jaccard distance, are used widely in a variety of different fields, including complex networks [4].

Because of this widespread use, lower variance and availability of pre-built computational functions, we recommend the Jaccard distance as a vertex-to-vertex distance measure for graph clustering. For example, the NetworkX library offers a Jaccard coefficient function, which we use in this work [11].

6 Metric Space and the Jaccard Distance

A metric space is a set of points that share a distance function. This function must have the following three properties:

$$g(x, y) = 0 \Leftrightarrow x = y \tag{1}$$
$$g(x, y) = g(y, x) \tag{2}$$
$$g(x, z) \leq g(x, y) + g(y, z)\,. \tag{3}$$

In the case of the Jaccard distance, the first two properties are immediately apparent. They are direct consequences of the definitions of set operations. The third property, the triangle inequality, was shown to hold by Levandowsky and Winter [13, 4].

7 Conclusion

We show that Jaccard and Otsuka-Ochiai distances, when averaged over clusters, very accurately follow the evolution of intra-cluster density. They are both shown to vary in an opposite direction to intra-cluster density. This variation has been observed to be robust to noise from inter-cluster edge probability and variations in cluster sizes. Finally, we also show that Jaccard distance displays lower variance than Otsuka-Ochiai distance.

Our future work will focus on a study of these distances on weighted graphs. We also intend to conduct empirical comparisons to commute and amplified commute distances. We are interested in studying the statistical properties of all these distances when averaged over clusters.

Acknowledgements

- The work of Alexander Ponomarenko was conducted within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE).
- The authors wish to thank the organizers of the The 10th International Conference on Network Analysis at the Laboratory of Algorithms and Technologies for Networks Analysis in Nizhny Novgorod.

References

1. M. Aramon, G. Rosenberg, E. Valiante, T. Miyazawa, H. Tamura, and H.G. Katzgraber. Physics-Inspired Optimization for Quadratic Unconstrained Problems Using a Digital Annealer. *Frontiers in Physics*, 7, Apr 2019.
2. Christian Bauckhage, Nico Piatkowski, Rafet Sifa, Dirk Hecker, and Stefan Wrobel. A QUBO formulation of the k-medoids problem. In Robert Jäschke and Matthias Weidlich, editors, *Proceedings of the Conference on "Lernen, Wissen, Daten, Analysen", Berlin, Germany, September 30 - October 2, 2019*, volume 2454 of *CEUR Workshop Proceedings*, pages 54–63. CEUR-WS.org, 2019.

3. R.S. Burt. Positions in Networks*. *Social Forces*, 55(1):93–122, 09 1976.
4. E. Camby and G. Caporossi. The extended Jaccard distance in complex networks. *Les Cahiers du GERAD*, G-2017-77, 09 2017.
5. N. Fan and P.M. Pardalos. Linear and Quadratic Programming Approaches for the General Graph Partitioning Problem. *J. of Global Optimization*, 48(1):57–71, September 2010.
6. N. Fan and P.M. Pardalos. Robust Optimization of Graph Partitioning and Critical Node Detection in Analyzing Networks. In *Proceedings of the 4th International Conference on Combinatorial Optimization and Applications - Volume Part I*, COCOA'10, pages 170–183, Berlin, Heidelberg, 2010. Springer-Verlag.
7. S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, February 2010.
8. S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.
9. Y. Fu and P.W. Anderson. Application of statistical mechanics to NP-complete problems in combinatorial optimisation. *Journal of Physics A: Mathematical and General*, 19(9):1605–1620, June 1986.
10. F. Glover, G. Kochenberger, and Y. Du. A Tutorial on Formulating and Using QUBO Models. *arXiv e-prints*, page arXiv:1811.11538, June 2018.
11. A.A. Hagberg, D.A. Schult, and P.J. Swart. Exploring Network Structure, Dynamics, and Function using NetworkX. In G. Varoquaux, T. Vaught, and J. Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.
12. P. Jaccard. Étude de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 01 1901.
13. M. Levandowsky and D. Winter. Distance between Sets. *Nature*, 234, November 1971.
14. A. Lucas. Ising formulations of many NP problems. *Frontiers in Physics*, 2:5, Feb 2014.
15. P. Miasnikof, A.Y. Shestopaloff, A.J. Bonner, and Y. Lawryshyn. *A Statistical Performance Analysis of Graph Clustering Algorithms*, chapter 11. Lecture Notes in Computer Science. Springer Nature, 6 2018.
16. P. Miasnikof, A.Y. Shestopaloff, A.J. Bonner, Y. Lawryshyn, and P.M. Pardalos. A density-based statistical analysis of graph clustering algorithm performance. *Journal of Complex Networks*, 8(3), 08 2020. cnaa012.
17. A. Ochiai. Zoogeographical studies on the soleoid fishes found in japan and its neighbouring regions-i. *NIPPON SUISAN GAKKAISHI*, 22(9):522–525, 1957.
18. A. Ponomarenko, L. S. Pitsoulis, and M. Shamshetdinov. Overlapping community detection in networks based on link partitioning and partitioning around medoids. *CoRR*, abs/1907.08731, 2019.
19. L. Ostroumova Prokhorenkova, P. Prałat, and A. Raigorodskii. Modularity of Complex Networks Models. In A. Bonato, F.C. Graham, and P. Prałat, editors, *Algorithms and Models for the Web Graph*, pages 115–126, Cham, 2016. Springer International Publishing.
20. L. Ostroumova Prokhorenkova, P. Prałat, and A. Raigorodskii. Modularity in several random graph models. *Electronic Notes in Discrete Mathematics*, 61:947 – 953, 2017. The European Conference on Combinatorics, Graph Theory and Applications (EUROCOMB'17).
21. S.E. Schaeffer. Survey: Graph clustering. *Comput. Sci. Rev.*, 1(1):27–64, August 2007.

22. U. von Luxburg, A. Radl, and M. Hein. Getting lost in space: Large sample analysis of the resistance distance. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2622–2630. 2010.
23. U. von Luxburg, A. Radl, and M. Hein. Hitting and commute times in large random neighborhood graphs. *Journal of Machine Learning Research*, 15(52):1751–1798, 2014.