Performance traits of a newly proposed modularity function for spatial networks: Better assessment of clustering for unsupervised learning

Raj Kishore1*, S Swayamjyoti1, Zohar Nussinov2 and Kisor K Sahu1

1School of Minerals, Metallurgical and Materials Engineering, Indian Institute of Technology Bhubaneswar- 752050, India
2Department of Physics, Washington University in Saint Louis, MO-63130-4899, USA

E-mail : *rk16@iitbbs.ac.in

Abstract: The “best” partition of a given network helps in revealing its naturally identifiable structures. The most modular structure is often considered as the best partition. Modularity function, is an objective measure of the quality of partitioning in a given network with that of a random graph (“Null model”), where edge between any two nodes is equally probable, are inappropriate to use for spatially embedded networks. Earlier we have proposed a new modularity function, which does not compare the network with a null model. We have analyzed a 2D and 3D granular networks which can be considered as a spatially embedded network. In all considered systems new method identifies the better partition. New function properly detects the better modular partition in 2D as well as in 3D granular assemblies as compared to the most commonly used modularity function, known as Newman modularity function, and thus is more suitable for unsupervised machine learning.

Keywords: Modularity, network partitioning, granular network, unsupervised machine learning

1. INTRODUCTION

Machine learning (ML) is changing the face of the technology and thereby almost every aspect of human civilization at a frenetic and feverish pace. ML comes in two basic flavors: viz. supervised and unsupervised learning (though one might count semi-supervised and reinforcement methods as additional classes). The former method is presently grabbing most of the attention because of the recent breakthroughs in the development of computational paradigms such as convolutional neural networks, deep learning and also in the painstaking construction of training data repositories, such as “Imagenet”[1] etc. For ML to realize its true potential, it must be applicable to generic systems, going beyond few restricted classes for which large training data is available. Therefore, unsupervised learning paradigm needs serious attention, since large training data repositories can only be constructed for finite (if not few) number of object types, out of countably infinite types that exist in natural world. One of the most important approach to unsupervised learning is clustering, wherein the problem is first transformed to an equivalent graph, following which one aims to put relatively similar objects in same group and thereby achieving a near-optimal partitioning. It is imperative to find a method to convert the subjective notion of “quality of partitioning” to an objective and quantifiable parameter. The most popular choice forth is “Modularity”. Modularity uses a reference random graph, called “null model” which has identical number of vertices and edges with that of the original graph. The connectivity between the nodes are randomized, but the degrees of respective vertices are preserved [2]. The quality of partition is benchmarked against this null model. It is abundantly clear
that, there exists many different choices for the null model and their suitability depends on the objective of the partition as well as the intrinsic structure of the network under study [3-4]. Here we focus only on static spatial networks. They are primarily characterized by the strength of interaction between its vertices (entities), which are embedded in a metric space. The interaction between entities in such a system determines its features, which manifests in the form of network hubs [5]. Absence of large and obvious hubs makes the network relatively feature-free and causes difficulty for many partitioning algorithms. In the present study, network with and without such obvious hubs are examined. These networks are created from ensembles of 2D and 3D granular particles. These networks are created by considering particle centers as node. Edge between any two particles is defined by their physical contacts. For further simplification, we assume all edges to be identical (we do not make any distinction between the edges by the sizes of vertices that it is connected to). Granular networks belong to a particular class of physically embedded networks where, geometrical constraints (neighborhood condition) preclude the possibility of existence of long distance edges. To generate visually distinct node hubs (high degree nodes), packing of different size particles is used. Quality assessment of the partitioning is mostly performed by calculating modularity of the network. Among many different modularity functions [6-7], Newman-Girvan (NG)modularity [8] is most commonly used for real networks. The NG modularity \( Q_{NG} \) is given as

\[
Q_{NG} = \frac{1}{2m} \sum_{i,j} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(\sigma_i, \sigma_j) 
\]

(1)

where, \( m \) is the edge count in the network, \( \sigma \) denotes community membership, \( k_i \) and \( k_j \) are the strength of \( i^{th} \) and \( j^{th} \) node respectively. \( A_{ij} \) is the connectivity matrix (\( A_{ij}=1 \), if an edge exist between node \( i \) and \( j \), and zero otherwise). \( \delta \) is the Kronecker delta and have unit value if node \( i \) and \( j \) belongs to same community and zero otherwise. NGmodularity compares the number of actual connections (\( A_{ij} \)) in a network to that of a null model (\( k_i k_j / 2m \)), therefore, it is more appropriate for the networks where the probability of an edge between any two arbitrary nodes is the same. However in granular networks, or more generally for networks embedded in a metric space, edges between nodes are typically confined within physically interacting neighborhood. NG function penalizes for missing edges (\( A_{ij}=0 \)) between nodes belonging to the same community even if an edge between them is physically not possible. Based on these considerations, a new modularity function, inspired by the NG method but modified to address these issues, was proposed. This formulation is extremely simple and readily applicable in unsupervised learning methods and is discussed at lengths in Section 2.4. It is demonstrated that the best partition identified by the modified modularity over wide range of resolution closely matches with the naturally identifiable structures.

2. METHODS

2.1 Creation of particulate ensembles: Realistic simulation of packing by Discrete Element Method (DEM)

In this study, Discrete Element Modelling (DEM) is used to generate two dimensional granular network, since it closely mimics the reality and is computationally the most attractive choice to model such networks. DEM was first proposed by Cundall and Strack [9-10]. It is used to simulate the dynamics of ‘soft’ particles in the present study. The detailed information of packing algorithm which includes information on contact models and the numerical scheme used in the present DEM simulation can be found in [11]. This method is used to generate the granular assembly of 5298 different size particles (disks in 2D) ranging from 0.01 to 0.1m as shown in Figure 1 (a). Initially, all these disks are generated using random position generator inside a circular box with initially low density (~30%). The dense packing of particles is obtained using a hypothetical centripetal force acting towards the center of the box [12]. This centripetal force is the sole externally applied force. Due to this force all the particles moves towards the center of box and finally a dense packed structure
is formed after numerous collision among themselves.

2.2 Construction of granular network

Two classes of granular networks are studied in this article. Class 1 is a two-dimensional (2-D) granular ensembles modelled by DEM (as discussed in Section 2.1). The given granular ensemble is converted into network by assuming particles centre as node and edge is considered between these particles, if they are physically interacting (the distance between their centres is less than or equal to the sum of their radii) with each other. The final network is depicted in Figure 1(b).

Class 2 is a three-dimensional (3-D) granular ensembles obtained by packing similar size atoms in a cubic lattice. This ensemble is obtained after manual removal of all the particles from the central XZ and YZ plane (one atomic layer) of the box of atom to create distinct boundary. Therefore the structural feature of this packing is distinctly clear: four unconnected small columnar packings separated by two orthogonal boundary planes as is clear from Figure 1(c). Corresponding network can be found in Figure 1(d). In this case DEM was deliberately not used as it might lead to non-crystalline packing and the visual clarity of 2-D projection of such a packing onto a piece of paper might not be as perceptive as that of a crystalline packing. This visual comparison is a crucial requirement for making a fair comparison between the two modularity functions since developing an objective measure is a non-trivial job and it’s interpretation might be even harder.

Figure 1. Different granular assemblies and their corresponding networks. (a) Assembly of different size of circular particles (from 0.01 to 0.1 m radius) and (b) its corresponding granular network. Colorbar in (a) represents the coordination number of particles. (c) 3-D cubic crystal structure, where all the particles from the central XY and XZ planes are removed and (d) its corresponding granular network.

2.3 Finding the communities by Rhonhoyde and Nussinov model

The best partition of a network should consists of groups of densely connected nodes termed as “communities” that are sparsely connected to other communities. The efficiency and accuracy of a
graph partitioning model is mainly affected by the ‘quality function’ it uses, and its implementation. The quality function should not only appreciate the connected edges but should also restrict large number of missing edges in a community. In the present work, network partitioning has been done by using spin-glass-type Potts model algorithm, developed by Peter Ronhovde and Zohar Nussinov (RN model) [13]. The spin interaction is modelled by Potts model and optimal partitioning is associated with minimization of system energy (Hamiltonian). For a detailed description, please see [13] and the computer code can be found in (file: RN.RAR available at https://sites.google.com/a/iitbbs.ac.in/kks-research-work/research-data). This model is highly accurate, a local model for general graphs (weighted, unweighted, and directed), and use of the structural resolution parameter $\gamma$, that makes it free from the resolution limit. The ground state of the potts model Hamiltonian facilitates objects of similar spins in same state and vice versa.

$$H(\{\sigma\}) = -\frac{1}{2} \sum_{i,j} (a_{ij}\sigma_i - b_{ij}\sigma_j)\delta(\sigma_i, \sigma_j)$$

(2)

where, $H$ is the Potts Hamiltonian of the system, $a$ and $b$ are the edge weights for connected and missing edges respectively. $A$ is the connectivity matrix, $J_{ij} = I - A_{ij}$. The structural resolution parameter, $\gamma$ issued to adjust the resolution of the community solution, $\sigma$ and $\delta$ have same meaning as in Eq. (1).

2.4 Modified Modularity function

We have proposed a new modified modularity function [14] which can be used for clustering the sparse networks (for e.g. granular networks studied here) [15]. The modularity function is given as:

$$Q(\sigma) = \frac{1}{2m} \sum_{i,j} (a_{ij}\sigma_i - b_{ij}\sigma_j)\delta(\Delta x_{ij})\theta(\Delta x_{ij}) \left(2\delta(\sigma_i, \sigma_j) - 1\right)$$

(3)

Here $a_{ij}$ and $b_{ij}$ are the strength (Note: it is different from edge weights used in RN method in Eq. (2)) of connected and missing edges between nodes $i$ and $j$ respectively. Edge strengths, $a_{ij}$ and $b_{ij}$, are the difference between the local node strength of node $i$ and $j$ and the average node strength $<k>$ of the network,

$$a_{ij} = b_{ij} = \left(\frac{k_i + k_j}{2} - <k>\right)$$

(4)

where, $<k> = \frac{1}{N} \sum_{r=1}^{N} k_r$, and $N$ is total number of nodes and $k$ is the node degree.

$$\{2\delta(\sigma_i, \sigma_j) - 1\} = \begin{cases} 1 & \sigma_i = \sigma_j \ (i.e \ in \ same \ community) \\ -1 & \sigma_i \neq \sigma_j \ (i.e \ in \ different \ communities) \end{cases}$$

(5)

Our method has mainly two distinctions over other existing methods of modularity calculation. First, it is independent of any null model comparison. Second, it takes into accounts the effect of inter-community edges while determining the most modular structure. New function has also restricted the over penalization for missing links between two distant nodes within same community by using Heaviside unit step function $\theta(\Delta x_{ij})$. It defines the neighborhood and penalizes only for the missing links inside the neighborhood. Here $\Delta x_{ij} = x_i - x_j$ is the difference in Euclidian distance between nodes $i, j$ and a predefined cutoff distance for neighborhood $x_c$ which is chosen as $x_c = 1.05(R_e + R_i)$ for the present study (where, $R_e$ is the radius of particle).
This proposed new modularity function is based on the difference between local node strength and the average node strength of the network (equation 4). A highly linked community structure will increase the function’s value as it represent a structure where average strength of communities will be higher than the average strength of the network.

3. RESULTS

3.1 Results for Class-1: 2-D packing

The modularity for different partition is calculated both by NG function and the new function (Figure 2a). The best partition suggested by the new function (Figure 2b) has high visual correlation with the network as it is separating most of the network hubs into distinct clusters. Whereas the partition given by NG method (Figure 2c) has no such immediate visual correlation as it places many hubs in one cluster. In order to make the comparison more compelling, we have selected the first quadrant of the network and have presented the community structures for this portion by using the new modularity function in the top panel of Figure 2(d). In the bottom panel of the same plot, we have selected the same portion (first quadrant of the network) presented in the community structures by the NG modularity. It is extremely difficult to miss the very high visual correlation with that of the original network and the community structure obtained by the new modularity function. It is therefore reasonable to argue that the perception of modularity by the new function is more “human-like” as compared to that of NG method. This is one of the major objectives for “machine-vision” and therefore, if this trend turns out to be more generic, it may become more suitable for, at least, very low level unsupervised learning.

Figure 2. Comparison of best partition obtained using NG modularity vs. new modularity. (a) Modularity of partition obtained at different γ value using RN model (b) best partition depicted from new modularity function (at γ=0.5) (c) best partition given by NG method (at γ=0.004) (d) First quadrant of the network and its corresponding best partition suggested by NG and new function
3.2 Results for Class-2: 3-D packing

In the previous assembly which was 2D structures, we have shown that new modularity function has provided better results in picking the most modular partition. But it’s accuracy for 3D network is still not known. So, in order to examine its performance for 3D structures, we have studied two different 3D assemblies. In this class we have developed a cubic crystal inside a cuboidal box and have manually removed all the particles of one XZ and YZ plane at the center of the box to create distinct boundaries. This arrangement makes this assembly more featured with clear distinction between four small cuboidal structures (Figure 3a). These four structures are totally unconnected and thus, in best partition of this network they should be considered as separate entities. We have partitioned the 3D networks using RN method at different $\gamma$ values and have calculated the modularity of each partition using both the modularity functions (Figure 3b). The modularity curves for both the functions shows similar trend in the region $10^3 > \gamma > 7 \times 10^{-4}$ (from (2) to further right in (b)) and beyond this, they started behaving differently. Interestingly, in the region $2 \times 10^{-5} > \gamma > 7 \times 10^{-4}$ (shaded region in (b)), the values of both the function are constant. However, for the new modularity function it was a peak, and for NG function, it was a local minimum. As the best partition suggested by both the functions is different, a visual inspection is needed in order to compare their accuracy. Since in 3D networks, visualization of the community inside the network is difficult, we have selected different XY planes at different heights (Z positions, as indicated by red arrows in Figure 3a). We have plotted the position of the selected planes (disks in 2D) of the colored based on their coordination number (panel (1) in Figures 3(c-f)). The coordination number of the particles residing in these XY planes varies from 3 to 6 (as maximum neighbors in a cubic system can be 6). We have also colored the particle positions based on their community membership in the best partition as suggested by NG function, and new modularity function. The best partition given by NG method has many lower-mode ordering (encircled; which is clearly not the best solution, (panel (2) in Figures 3(c-f)), whereas maximum modular community structure detected by new function (Figure 3c(3)-3f(3)) is most stable in the range of $7 \times 10^{-4} < \gamma < 2 \times 10^{-5}$ and it is clearly picking the four small cuboidal structures as separate communities. Therefore it is clear that, even in its peak, NG modularity is selecting mixed modes, whereas the new function, in the given range of $\gamma$, is picking up only the pure primary mode without any false partitioning of the natural structure (four small cuboidal blocks) of the network. We have also analyzed a partition at $\gamma=1 \times 10^{-5}$ (panel (4) in Figures 3(c-f)) as at this resolution both the functions are showing opposite modularity response: NG is increasing whereas modified modularity is decreasing. NG modularity ranks the partition as: (2) > (4) > (3) whereas new modularity function predicts the following ranking (3) > (4) > (2). Through visual inspection of community structure in different planes selected, structure at (2) has many lower-mode ordering (encircled in panel (2) in Figures 3(c-f)), which is lower in (4) and totally absent in the community structure at (3). The logical perception of modularity decrement should follow the order (3) > (4) > (2), as predicted by new modularity function. So it is reasonable to conclude that new modularity function is more efficient than NG function in picking the “right” partition even in 3D assemblies.
Figure 3. (a) 3D assembly of cubic crystal structure with 100% particles removed from one XZ and YZ planes at the center to create disordered regions. (a*) is the top view (2D projection). (b) Modularity of the partition obtained at different $\gamma$ value using RN model and (c-f) XY planes at different Z positions as identified in (a).

In all (c-f), panel (1) represents the coordination number with colorbar next to it identify its value; panels (2-4) corresponds to the community partitions identified in (b) and depicted by different colors. It is clear that in all (c-f), (3) represents the best partition as predicted by algorithm. NG method predicted that (2) and (4) are better partition than (3), which is clearly not the case. Therefore, new modularity clearly outperforms the NG method for this class.

4. CONCLUSIONS

The summary of the entire study is presented in Table 1, which clearly tells that in both 2-D and 3-D granular networks, the new modularity function performs better than the NG modularity function.
Table 1: Summary of results for all the different classes of network studied in this work. It clearly indicates that the new modularity function is superior to the NG modularity for the networks studied in this article.

| Tag   | Particle type          | Packing method | Network preparation protocol | Community detection | Best partition finding          | Best performance          |
|-------|------------------------|----------------|-------------------------------|---------------------|--------------------------------|--------------------------|
| Class-1: 2-D | Different size disk | DEM            | Densely packed network        | RN METHOD           | NG vs. modified modularity     | Modified modularity      |
| Class-2: 3-D | Same size spheres | Unit cell repetition | 3-D cubic network; spheres in ordered pattern | RN Method           | NG vs. modified modularity     | Modified modularity      |

Here we have tested the performance of a newly proposed modularity function on different spatially embedded 2D and 3D networks. The new modularity function is convincingly picking the correct “best” resolution scale at which the community structure has highest alignment with the natural feature of the structures. It is capable of detecting node hubs if present. It is also equally good for feature-free networks.

REFERENCES:

[1]. Krizhevsky, A., Sutskever, I., & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105) (2012).
[2]. Newman, M. E. J., & Girvan, M. Finding and evaluating community structure in networks. Phys. Rev. E, 69(2), 026113 (2004).
[3]. Reichardt, J., & Bornholdt, S. Statistical mechanics of community detection. Phys. Rev. E, 74(1), 016110 (2006).
[4]. Kishore, R., Swayamjyoti, S., Das, S., Gogineni, A. K., Nussinov, Z., Solenov, D., &Sahu, K. K. (2020). Visual Machine Learning: Insight through Eigenvectors, Chladni patterns and community detection in 2D particulate structures. arXiv preprint arXiv:2001.00345.
[5]. Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. Nature reviews neuroscience, 10(3), 186-198.
[6]. Good, B. H., De Montjoye, Y. A., & Clauset, A. (2010). Performance of modularity maximization in practical contexts. Physical Review E, 81(4), 046106.
[7]. Duch, J., & Arenas, A. (2005). Community detection in complex networks using extremal optimization. Physical review E, 72(2), 027104.
[8]. Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. Physical review E, 69(6), 066133.
[9]. Cundall, P. A., & Strack, O. D. (1979). A discrete numerical model for granular assemblies. Geotechnique, 29(1), 47-65.
[10]. Cundall, P. A. (1988, June). Formulation of a three-dimensional distinct element model—Part I. A scheme to detect and represent contacts in a system composed of many polyhedral blocks. In International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts (Vol. 25, No. 3, pp. 107-116). Pergamon.
[11]. Kishore, R., Das, S., Nussinov, Z., & Sahu, K. K. (2016). Kinetic instability, symmetry breaking and role of geometric constraints on the upper bounds of disorder in two dimensional packings. Scientific reports, 6, 26968.
[12]. Kishore, R., Krishnan, R., Satpathy, M., Nussinov, Z., & Sahu, K. K. Abstraction of meso-scale network architecture in granular ensembles using ‘big data analytics’ tools. J. Phys. Comm., 2(3), 031004 (2018).

[13]. Ronhovde, P., & Nussinov, Z. Local resolution-limit-free Potts model for community detection. Phys. Rev. E, 81(4), 046114 (2010).

[14]. Kishore, R., Gogineni, A. K., Nussinov, Z., & Sahu, K. K. (2019). A nature inspired modularity function for unsupervised learning involving spatially embedded networks. Scientific reports, 9(1), 1-9.

[15]. Kishore, R., Satpathy, M., Parida, D. K., Nussinov, Z., & Sahu, K. K. (2019, September). Detection of Lung Nodules Using Unsupervised Machine Learning Method. In International Conference On Computational Vision and Bio Inspired Computing (pp. 463-471). Springer, Cham.