Network Analysis of Saccharomyces Cerevisiae Colony: Relation between Spatial Position and Generation

Widya Meiriska¹, F A Purnama¹, D P P Aji², D Aprianti², and S Viridi¹

¹ Nuclear and Biophysics Research Division, Physics Department, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Jl. Ganesha 10, Bandung 40132, Indonesia
² Computational Science Department, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Jl. Ganesha 10, Bandung 40132, Indonesia

Email: widya.meiriska@gmail.com

Abstract. Yeast colony growth which can grow and produce buds can be modelled by using circular granular cells in two-dimensional form. The organizations of particles in granular materials as modelled as yeast colony growth have complex organizations on a spatial scale. Such organizations may affect how materials respond or reconfigure when exposed to external interference or loading. Particles have been studied in theoretical and require the development and application of appropriate mathematical, statistical, physical and computational frameworks. Usually, granular materials have been explored using particles or circuit models that are implicit. Today the development of network science has emerged as a powerful approach to investigate and characterize heterogeneous architecture in complex systems and diverse methods have yielded interesting insights into granular materials. In this study will learn granular materials (yeast colony growth) with the network-based approach and explore the potential of these frames to provide a useful description of yeast colony growth. This study will focus on finding the relation between the spatial position of the cells and generation that is formed.

1. Introduction

Nowadays, a sophisticated methodology known as a network-based approach has been developed to translate physics into mathematical models and computations so can be analysed and understood to provide a predictive understanding of object behaviour. It turns out that many complicated systems can be explained by this network-based approach. The development of network science has emerged as a strong approach to investigate and characterize heterogeneous architecture in complex systems. In this study, we will try to apply a network-based approach giving information and description which can improve our understanding of the two-dimensional granular system. The granular system which is reviewed is the budding yeast colony growth that has been simulated previously. The simulation is performed to mimic the colony formation. The initial configuration of the colony in the simulation is set similar to the initial configuration of the colony from the observation video of real colony formation [1].

2. Graph Theory

In mathematics, networks are known as graphs. Graph is a set of objects called vertices or nodes connected by edges. Graph is used to represent discrete objects and relationships between these objects.
A graph $G$ can be expressed as $G = (V; E)$. Graph $G$ consists of a number of $V$ that contain nodes and a number of $E$ containing the sides of the graph. The set $E$ is expressed as a pair of nodes in $V$. The development of algorithms to handle graphs will have a major impact on computer science [2]. A structure graph can be developed by giving weight to each side. Weighted graphs can be used to symbolize many different concepts. Another extension to the graph is to make the sides point, which is technically called a directed graph.

In addition to using the set, the network can also be represented in the form of a matrix to make it easier for someone to analyse a graph if calculations are needed in the analysis process. This interpretation in the matrix form is known as an adjacency matrix. If each node is labelled it can be shown side as a pair $(i, j)$, so that it can be expressed using an adjacency matrix which can be written with element $A_{ij}$. The adjacency matrix is the most commonly used in the graph representation. This matrix only contains 0 and 1, then the matrix is also called a zero-one matrix [3]. The adjacency matrix can be written as follows:

$$A = \begin{bmatrix}
0 & 1 & 1 & 0 \\
1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 \\
0 & 1 & 1 & 0
\end{bmatrix}$$

$A_{ij} = \begin{cases}
1, & \text{if there is an edge between } i \text{ and } j \\
0, & \text{if there is no edge between } i \text{ and } j
\end{cases}$

### 2.1 Network Parameters

There are many parameters related to a graph. By knowing these parameter values can provide information about the graph.

#### 2.1.1 Degree

The degree is the number of sides that is associated with the node and is defined by its adjacency matrix. In undirected graph, the degree can be calculated using the following equation:

$$k_i = \sum_{j=1}^{N} A_{ij} = \sum_{j=1}^{N} A_{ji}$$

In a directed graph, the degree of the node is divided into two components, namely the number of sides that come out from the node (out-degree) and the number of sides that enter the node (in-degree).

$$k_i^{\text{out}} = \sum_{j=1}^{N} A_{ij}$$

$$k_i^{\text{in}} = \sum_{j=1}^{N} A_{ji}$$

So that the total degree of a node is defined as

$$k_i = k_i^{\text{out}} + k_i^{\text{in}}$$
Whereas in a weighted network, degrees are analogous to strength, namely the number of weights on the sides associated with the set of vertices.

2.1.2 Degree Centrality
Degree centrality shows nodes that have an important role in the network based on the number of relationships that they have with other nodes [4]. The degree centrality of each node can be formulated as follows:

\[
C_D(i) = \frac{\sum_{j=1}^{N} A_{ij}}{(N-1)(N-2)} \quad (i \neq j)
\] (5)

2.2 K-Means Clustering
K-Means clustering is a machine learning algorithm that is used to partition data into desired clusters. The K-Means method attempts to classify existing data into several groups, where data in one group has the same characteristics as each other and has different characteristics from the data in the other groups. In other words, this method seeks to minimize variations between data that is existed within a cluster and maximize variation with data in other clusters [5]. Distance space is used to calculate the distance between data and centroids that has been decided. The equation is as follows:

\[
d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}
\] (6)

3. Simulation
3.1 Network Data
This network analysis was applied to the data from the simulation of the budding yeast colony growth that had been simulated before.

In simulation, yeast cells which initially only one cell continues to increase as time goes on to reach 100 yeast cells. In this study some assumptions are given including:
- Yeast cell 1 is the first generation, born from the result of self-splitting which in network science can be said that cell number 1 doing self-loops in the initial conditions formed
- The network formed is a directed graph and network that has weight.
- The weight that given is the distance between two interconnected nodes which are then divided by the average diameter of the yeast cell that is connected, which uses the following equation:
\[ r_y = \frac{\sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}}{\frac{1}{2}(d_i + d_j)} \] (7)

### 3.2 Encoding Technique

Network analysis starts with representing data into the graph representation. This step is also known as an encoding technique. Encoding technique aims to represent the graph into an adjacency matrix. In this study, the type of network used is not the simple graph because it has self-loops and direction. Encoding is done by describing the data of the relationship between each node (yeast cells) and changing it into binary data (value 0 or 1). This encoding technique is used to calculate all network parameters in this study and visualize the structure of the yeast colony network.

#### Table 1. Simulation Algorithm.

| Algorithm |
|-----------|
| 1: Select the data that will be observed from the existing raw data |
| 2: Modify the selected data, so that it can be processed using Python 3.6 |
| 3: Determine the parameters that will be designated as nodes and edges |
| 4: Perform encoding techniques on selected nodes |
| 5: Calculate the network parameters |
| 6: Data visualization |
| 7: Using K-Means clustering algorithm for the data partitioning process |

### 4. Result and Discussion

After encoding the data, the visualization of the adjacency matrix will be obtained as shown below. The \(x\) and \(y\) axes indicate the selected node, the child cell (\(A\)) and the mother cell (\(I\)). After encoding techniques and getting the adjacency matrix, the degree centrality can be calculated. Here are some results of network parameter calculations. The following is the result of calculating the degree centrality’s value.

#### Table 2. The number of cells which has many descents and higher degree centrality’s value.

| Cell | Descent | Degree Centrality |
|------|--------|-------------------|
| 1.0  | 14     | 0.1515            |
| 5.0  | 9      | 0.101             |
| 7.0  | 9      | 0.101             |
| 2.0  | 7      | 0.081             |
| 4.0  | 7      | 0.081             |
| 17.0 | 6      | 0.071             |
| 26.0 | 6      | 0.071             |
| 3.0  | 5      | 0.061             |
| 10.0 | 5      | 0.061             |
| 6.0  | 4      | 0.051             |

#### Figure 4. Adjacency Matrix.
The value of the degree centrality provides information the node that has the most number of descents and becomes the central figure on the network. In the case of yeast colony growth, it means loading information on which cells are the most productive cells with the highest number of children.

![Figure 5. Budding yeast colony growth network structure](image)

(a) the directed graph with weight (b) cluster tree visualization.

In Figure 5 (a) is the structure of the yeast cell growth network, a clear picture of the most productive mother and the distance between mother and child is obtained. The type of network used is a directed graph, which shows the relationship between mother and child and a weighted graph that provides an overview of the distance between the mother and child cell. In Figure 5 (b) is other data visualization of yeast colony growth. In this visualization, cells can be seen along with the sequence of generations of cells. This visualization is able to describe the relationship between mother and child and the generation that is formed of the yeast cell.

Furthermore, we try to analyse the influence of distance on the generation. In the network structure mapping in Figure 5 has been given weight on the network that shows the distance between mother cell and child cells. Here is the distribution of distance from the network.

![Figure 6. Distance’s degree.](image)

| Distance | Degree |
|----------|--------|
| 0        | 1      |
| 1        | 61     |
| 2        | 20     |
| 3        | 7      |
| 4        | 7      |
| 6        | 1      |
| 7        | 2      |
| 9        | 1      |
The process of determining generation based on distance also uses a machine learning algorithm, namely K-Means clustering. In this process several assumptions are carried out: centroid in this algorithm shows the generation of networks. The network will be divided into four clusters (four centroids). The centroid value is determined based on the distance from the mother cell and child cell.

- Centroid 1 (mother) is a yeast cell that has distance 0 (self-loops)
- Centroid 2 (generation 1) is a yeast cell that has distance in the range of 4 - 9
- Centroid 3 (generation 2) is a yeast cell that has distance in the range of 3 - 2
- Centroid 4 (the last generation) is a yeast cell that has distance 1. This last generation is used for cells that no longer produce child.

After clustering, using this algorithm and assumptions above, the following result are obtained:

![Figure 7. The relation between distance and generation.](image)

The results of clustering based on this distance when compared with the results of the previous network structure have shown almost correct results, with an accuracy level of 87%. The errors occur here may be caused by the incorrect determination of centroid assumptions, and the tendency of cells that is not driven by new cell formation. Basically, the use of this clustering algorithm is still not able to provide truly reliable results (100% true), but it is enough to describe the conditions that occur in this yeast cell growth network.

5. Conclusion
Network analysis has been successfully carried out on the data of the yeast colony growth network and it is able to provide useful information including, the information about the most productive yeast cells based on the value of the degree centrality. Provides some visualizations yeast colony growth network. Moreover, has been succeeded in obtaining a relationship from the distance between the mother cell and child cell and the generation that is formed using K-Means clustering, which gives accuracy 87% true.

Acknowledgments
Authors say thank you to all those who have helped in this research and writing of this paper

References
[1] Devi Aprianti et al 2016 J. Phys.: Conf. Ser. 739 012026
[2] Newman M 2010 Networks: An Introduction (Oxford: Oxford University Press) chapter 6 pp 101-115
[3] Newman M 2003 Phys. Rev. E 67 026126
[4] Brandes U and Erlebach T 2004 Network Analysis: Methodological Foundation (Berlin:Springer) chapter 1 pp 12 - 20
[5] Smola A and Vishwanathan S 2008 Introduction to Machine Learning (Cambridge: Cambridge University Press) chapter 1 pp 32 - 34