Clustering for Items Distribution Network

M R Yudhanegara¹, S W Indratno², and R R K N Sari³

¹,²,³Statistics Research Group, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Jalan Ganesha 10 Bandung, 40132, Indonesia

Email: ¹mridwan.yudhanegara@s.itb.ac.id

Abstract. The distribution network can be represented as a simple graph. In this case, the delivery destination location is represented as a vertex, while the road passed is represented as an arc. The main concept of network clustering or graph clustering is that high connected vertices can be in the same cluster, while low connected vertices can be in different clusters. The method for network clustering in this paper is the Markov cluster. The purpose of items distribution network clustering i.e. to facilitate the company in item delivery based on road conditions.

1. Introduction
A network consists of a group of vertices are connected by an arc [1], hereinafter referred to as a connected graph. Network clustering is useful for grouping vertices of a graph into several clusters by considering the arc structure of the graph. So there are many arcs in a cluster and relatively few arcs between clusters. Graph clustering are important in finding the basic structure of a network.

Graph-based clustering is closely related to graph partitioning, where this method is used to find the optimal partition of a graph with any bound. On the graph partition, the required partition size is determined first. The purpose is to minimize some objective functions related to relationships that connect different partition elements. This is a fundamental difference between the partition graph and graph clustering, where in the cluster graph the discovery of groups is done naturally.

This paper discusses the distribution of items by clustering on an item distribution network. In the activity of items delivered, there have not been many cases it by clustering on the item distribution network. In this case, the network is represented in a simple graph.

The graph itself is a set of pairs \((V, A)\) with \(V\) is a finite and non-empty set of vertices \(V = \{v_0, v_1, ..., v_n\}\), and \(A\) is a set of pairs between two vertices called arcs, \(A = \{(v_i, v_j) : i, j \in V, i \neq j\}\). In this problem, the vertex represents as destination location while the arc represents a path that connects the location.

If vertex \(v_i\) and \(v_j\) are neighbors to each other, then vertex \(v_i\) or \(v_j\) and arcs \((v_i, v_j)\) are said to be incidents to each other [2]. The vertex \(v_i\) and \(v_j\) are adjacent if there is an arc between \(v_i\) and \(v_j\) or there is a pair \((v_i, v_j) \in A\) where \(v_i, v_j \in V\).

Graph \(G(V, A) = G\) is a weighted graph \(G\) on \(V\). The graph is a pair \((V, g)\), and \(g\) is a function that maps each pair of vertices \(V\) to real nonnegative numbers: \(g : V \times V \rightarrow \mathbb{R}^{\geq 0}\). Let \(G(V, g)\) be a weighted graph with the number of vertices \(|V| = n\). The association matrix of \(G\) on \(\mathbb{R}^{n \times n}_{\geq 0}\), denoted \(M_G\), is formed
by setting the entry \((M_G)_{i,j}\) equal to \(g(v_i, v_j)\). So that \(M_G \in \mathbb{R}^{n \times n}\) is the association matrix of graph \(G(V, g)\) where \(|V| = n\) and \(g(v_i, v_j) = (M_G)_{i,j}\).

The main concept of network clustering or graph clustering is that high connected vertices can be in one cluster, while low connected vertices can be in different clusters. As an illustration, the number of paths \((v_i, v_j)\) with the length \(h\) unit of length is greater if \(v_i\) and \(v_j\) are in the same dense cluster, and smaller if what is in a different clusters.

In a graph, random walk can occur that is a sequence of vertices generated from the starting vertex by selecting a path, passing to a new vertex and repeating the process. A random walk on the graph will produce a cluster until the vertices have been visited. Random walk can be started from any vertex, but in the case of distributing of items the starting vertex has been determined, namely depot \((v_0)\). Suppose that it starts from vertex \(v_0\), if the random walk will reach vertex \(v_i\) with high probability, then \(v_0\) and \(v_i\) must be grouped together.

Based on a graph, there are many links in a cluster and fewer relationships between clusters. If it starts from a vertex, and then does a random walk to the connected vertex, it is more likely to stay inside a cluster rather than walk between the two. This is the basis of the Markov cluster, which will then be used as a method for graphs clustering as items distribution network.

Random walk on graphs can be used to determine the existence of a group that is characterized by a tendency to gather a flow. In addition, random walk on graphs are calculated using the Markov chain. The Markov chain is a line of variables given the current state, while the previous state and the next are mutually independent. The probability for the next time step depends only on current probabilities [3]. The random walk is an example of a Markov chain using a probability transition matrix.

To complete network clustering, the procedure is carried out, namely by the Markov chain process through a random walk. This idea was first introduced by van Dongen [4], hereinafter referred to as the Markov cluster method. The idea of the Markov cluster method is to simulate flow in a graph, which consists of two purposes. The first goal is to strengthen the current strong flow, and the second goal is to weaken the current weak flow.

Not only van Dongen [4] who examines graph clustering. In addition, there is Zhu et al. [5] who examined graph clustering using the gradient flow method. But in this paper the material refers to van Dongen [4]. Because the items distribution network contains a strong and weak flow, which is the weight of each arc.

As information for readers, what is not discussed in this paper is about the rule-k-means algorithm. So, before clustering for item distribution network is carried out, first clustering for items of delivery is done based on its features.

In this case, the purpose of item distribution network clustering is to map items on similar delivery vehicles (large box cars, small box cars, motorcycles). Furthermore, exposure to the methods used will be presented in section 2 (Methods) after the introduction. For simulation data processing using the Markov cluster is presented in section 3 (Result and Discussion). The conclusions of this paper are presented in section 4 (Conclusion).

2. Methods

Before the graph is made as a representation of the destination location and the path that is passed, first the data are clustered using the rule-k-means method. The algorithm is an extension of the k-means algorithm used for clustering items based on numerical features and categorical features contained in the items. The basic form of the k-means algorithm is described in MacQueen [6]; Anderberg [7]; Everitt et al. [8]. The purpose is to map items into item delivery vehicles (large box cars, small box cars, motorcycles).

To quantify categorical data, the procedure is carried out by weighting. If the data item have categorical features, then \(X_i = (x_{i,1}, x_{i,2}, ..., x_{i,a})\) where

\[
x_{i,j} = w_{i,j}y_{i,j}, 1 \leq i \leq n
\]
and
\[ \sum_{i=1}^{n} w_i = 1, \]  
(2)
where \( y \) is categorical feature, and \( w \) is the weight.

The next procedure is weighting adjusted data by referring to Equation (1) and Equation (2), so it is obtained
\[ w_i = \begin{cases} 0, & \text{if } y_i = 0 \\ \omega_i, & \text{if } y_i \neq 0 \end{cases} \]  
(3)
where
\[ \omega_i = \frac{\sum_{j=1}^{n} (D_i y_j)}{D_i y_i} \]  
(4)

In the Markov cluster method, the flow simulation at the random walk is the core concept of finding clusters. The process of flow is easier in dense areas. The arc weight will be higher in the chain inside the cluster, and lower between the clusters.

There are two operations in the Markov cluster algorithm, namely expansion and inflation. Expansion is an operation to find the power of the Markov chain transition matrix, in this case the association matrix \( M_G + I_n \), with \( I_n \) is the identity matrix \( M_G \) of size \( n \times n \). Then the power numbers are called the expansion parameter \((e)\).

Inflation is an operation that is responsible for strengthening and weakening a condition. The inflation parameter \((r)\) controls, strengthening and weakening. Furthermore, inflation is defined according to van Dongen [4] as follows:

**Definition:**
Given matrix \( M_G + I_n = M \in \mathbb{R}^{n \times n}, M \geq 0 \) and a real nonnegative numbers \( r \). The matrix resulting from scaling each of the columns of \( M \) with power coefficient \( r \) is called \( \Gamma_r M \), and \( \Gamma_r \) is called the inflation operator with power coefficient \( r \). \( \Gamma_r : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n} \) is defined by Equation (5).

\[ (\Gamma_r M)_{i,j} = \frac{(M_{i,j})^r}{\sum_{k=1}^{n} (M_{k,j})^r} \]  
(5)

If the subscript is ommited, it is understood that the power coefficient equals 2.

The following are the steps of the Markov cluster algorithm according to Macropol [9]:

a. Input of graph to be clustered, parameter value of expansion and inflation.
b. Create an association matrix \( M_G \) whose columns and rows represent points on the graph.
c. Add loops, each of vertex, or in other words the association matrix \( M_G \) is added to the identity matrix \( I_n \).
d. Normalize the results matrix or create a transition probability matrix. At this stage, each entry of the matrix is divided by the sum of matrix entries in the column so that a matrix is obtained by the sum of entries from each column is 1.
e. Matrix expansion with the inflation parameter.
f. Matrix inflation with inflation parameters.
g. The clustering process is said to be convergent if there are no significant changes to the new cluster. If the condition has not been achieved, then repeat stages e and f until convergent conditions are obtained.
h. Matrix interpreters so that clusters of graphs are found.
Furthermore, the clustering process with the Markov cluster method is presented in the flow chart in Figure 1.

![Flow chart of Markov cluster method](image)

**Figure 1.** Flow chart of Markov cluster method for this study.

3. **Result and Discussion**

A summary of the data is presented by Table 1 with captions: object ($X_i$), distance of the object to the destination location (in kilometers) ($D_i$), weighting of road volume (in m$^3$) ($y_i$), weight ($w_i$), type of road ($x_i$), and Table 2.

**Table 1.** Data simulation for Markov cluster method in this study.

| $X_i$ | Weight (in gram) | Volume (in cm$^3$) | $D_i$ | Category            | $y_i$ | $D_i \times y_i$ | $w_i$ | $x_i$ |
|------|-----------------|-------------------|------|---------------------|------|-----------------|------|------|
| $X_1$ | 9000            | 165000            | 2    | Small way           | 1.6  | 3.2             | 0.000354687 | 0.000567499 |
| $X_2$ | 1500            | 3000              | 3    | Highway traffic     | 1.6  | 4.8             | 0.000532030 | 0.000851480 |
| $X_3$ | 1000            | 900               | 2    | Highway             | 20.7 | 41.8            | 0.000458875 | 0.094987281 |
| $X_{99}$ | 20              | 0                 | 3    | Highway             | 20.7 | 62.1            | 0.006883136 | 0.142480922 |
| $X_{100}$ | 500             | 250               | 4.5  | Main highway        | 48   | 216             | 0.023941344 | 1.149184498 |
| $X_{102}$ | 600             | 250               | 5    | Small way           | 1.6  | 8               | 0.000886716 | 0.001418746 |
Table 2. Data information records in Table 1.

| Features  | Type of features | Label         |
|-----------|-----------------|---------------|
| Weight    | Numerical       | -             |
| Volume    | Numerical       | -             |
| Type of road | Categorical     | Small way    |
|           |                 | Highway traffic |
|           |                 | Highway       |
|           |                 | Main highway  |

From the results of items clustering, groups of mapped objects are obtained, then shown in Table 3.

Table 3. Object groups on the type of delivery vehicle.

| Item delivery vehicles | Large box car | Small box car | Motorcycle |
|------------------------|---------------|---------------|------------|
| X₁₅                    | X₁₆           | X₈₂           | X₁         |
| X₁₀                    | X₅₃           | X₅₇           | X₂         |
| X₁₁                    | X₂₅           | X₈₈           | X₃         |
| X₂ₒ                    | X₅₅           | X₉₀           | X₄         |
| X₂₄                    | X₃₆           | X₉₈           | X₆         |
| X₂₉                    | X₃₇           | X₁₀₁          | X₇         |
| X₃₀                    | X₃₈           | X₃            | X₈         |
| X₈₆                    | X₃₉           | X₉            | X₉         |
| X₄₁                    | X₁₂           | X₃₄           | X₅         |
| X₅₃                    | X₁₃           | X₄₀           | X₇         |
| X₅₅                    | X₁₄           | X₄₂           | X₅₉        |
| X₆₈                    | X₁₅           | X₄₃           | X₆₀        |
| X₇₀                    | X₁₇           | X₄₄           | X₆₁        |
| X₇₃                    | X₁₈           | X₄₅           | X₆₂        |
| X₇₄                    | X₁₉           | X₄₆           | X₆₃        |

Based on Table 3, information is obtained that the large box car will deliver 8 item’s objects, 21 item’s objects will be delivered by small box cars, and 73 item’s objects will be delivered by motorcycle. Next is a graph based on the map of the items delivery vehicle. The graph with the object’s location in the big box car is depicted in Figure 2.(a).

Figure 2. (a) The location graph for large box car destination, and (b) the graph clusters.
The association matrix of the graph in Figure 2.(a) is represented in Figure 3.

![Association Matrix](image)

**Figure 3.** Association matrix ($M_G$) of Figure 2.(a).

Through the Markov cluster method with the expansion parameter $e = 2$ and the inflation parameter $r = 2$, the results of the destination location cluster are entries from the expansion operation and inflation convergent results to value 1 in the iteration 15th so that the number of clusters is 2. The first cluster \{v_5, v_{11}, v_{10}\}, and the second cluster \{v_{22}, v_{29}, v_{30}, v_{86}, v_{24}\}. The clusters in the form of networks are presented in Figure 2.(b). Then at the destination location the object in the small box car is depicted in Figure 4.

![Graph Clusters](image)

**Figure 4.** (a) The location graph for small box car destination, and (b) the graph clusters.

The association matrix of the graph in Figure 4.(a) is represented in Figure (5).
Figure 5. Association matrix \((M_G)\) of Figure 4.(a).

Through the Markov cluster method with the expansion parameter \(e = 2\) and the inflation parameter \(r = 2\), the results of the destination location clusters are entries from the matrix of expansion operations and inflation converging to value 1 in the iteration 16th so that the number of clusters is 4. The first cluster \(\{v_{38}, v_{23}, v_{37}, v_{35}, v_{36}, v_{39}, v_8, v_3\}\), second cluster \(\{v_{68}, v_{90}, v_{82}\}\) third cluster \(\{v_{55}, v_{70}, v_{41}, v_{87}, v_{16}, v_{25}\}\), then the fourth cluster \(\{v_{98}, v_{101}, v_{74}, v_{73}\}\). The clusters in the form of networks are presented in Figure 4.(b).

For the location of the object's destination for motorcycles are depicted in Figure 6.

![Figure 6](image)

Figure 6. (a) The location graph for the motorcycle destination, and (b) the graph clusters.

Through the Markov cluster method with the expansion parameter \(e = 2\) and the inflation parameter \(r = 2\), the result of the destination location cluster is the entry from the matrix of expansion operations and inflation converges to value 1 in the iteration 20th so that the number of clusters is 18. The cluster members are summarized in Table 4. Then for the cluster in the form of the network, it is shown in Figure 6.(b).
Table 4. Cluster member location for motorcycle destinations.

| Clusters | Members                                      |
|----------|----------------------------------------------|
| first    | v4, v77, v20, v19                           |
| second   | v31, v17, v49                               |
| third    | v92, v57, v45                               |
| fourth   | v44, v12                                    |
| fifth    | v96, v61, v65                               |
| sixth    | v21, v14, v1                                |
| seventh  | v46, v84, v66, v54, v76, v3                 |
| eighth   | v26, v80, v15, v6                           |
| ninth    | v106, v79, v95                              |
| tenth    | v7, v51, v43, v34, v48                      |
| eleventh | v91, v51, v71                               |
| twelfth  | v75, v9, v47                                |
| thirteenth | v78, v80, v81, v36, v99, v59, v60, v62, v63, v72, v97, v18 |
| fourteenth | v94, v28, v2                             |
| fifteenth | v8, v13, v40, v83, v54                        |
| sixteenth | v102, v89                                   |
| seventeenth | v69, v27                                   |
| eighteenth | v58, v33, v32, v67, v42, v52, v93            |

4. Conclusion
From the simulation results that have been done, it can be concluded that the clustering for items distribution network can facilitate companies in delivery of items, which can predict the delivery vehicle requirements that are in accordance with available road conditions. In addition, the items distribution network that includes a graph with an association matrix with a large size can be reduced by means of graph clustering.

In future works, we will work on clustering for distribution networks of dynamic items, where the arc weights which are the links between vertices in the network change every time.

5. Acknowledgment
The authors thank to the ITB’s Program of Research, Community Service and Innovation (P3MI ITB) 2019 for supporting funds.

6. References
[1] Taha H A 2016 Operations research: an introduction (10th edition) (New Jersey: Pearson)
[2] Diestel R 2017 Graph theory (5th edition) (New York: Springer)
[3] Ross S M 1996 Stochastic process (New York: John Wiley and Sons)
[4] van Dongen S 2000 Graph clustering by flow simulation (Utrecht: University of Utrecht)
[5] Zhu H, Chen C, Liao L, and Ng M K 2018 Multiple graphs clustering by gradient flow method, J. Franklin Institute 355 1819–1845
[6] MacQueen J 1976 Some methods for classification and analysis of multivariate observations Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1 281-297
[7] Anderberg M R 1973 Cluster analysis for applications (New York: Academic Press)
[8] Everitt B S, Landau S, Leese M, and Stahl D 2011 Cluster analysis (5th edition) (West Sussex: John Wiley and Sons)
[9] Macropol K 2009 Clustering on graph: the markov cluster algorithm (Utrecht: University of Utrecht)