Vol. 8, No. 4, August 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Comparative Analysis of Similarity-Based Edge Construction Methods for Village Welfare Index Networks

Abdullah Alhayad Arafah[1], Annisa[2], Sofyan Sjaf[3]

[1,2] School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia

[3]Department of Communication Science and Community Development, IPB University, Bogor, Indonesia

Received:
February 19, 2026
Revised:
June 20, 2026
Accepted:
July 23, 2026
Published:
August 22, 2026

Corresponding Author:
Author Name*:
Abdullah Alhayad Arafah
Email*:
alhayad.office@gmail.com

DOI:
10.63158/journalisi.v8i4.1769

Abstract. This study compares three similarity-based edge construction methods for village welfare networks from household welfare attributes in Data Desa Presisi. Households are represented as nodes, while edges denote computational similarity between household welfare profiles rather than social relationships. Five Village Welfare Index dimensions were discretized and transformed into binary one-hot representations. Pairwise similarities were calculated using Cosine Similarity, Jaccard Index, and Pearson Correlation, followed by automatic thresholding to retain strong relationships. Network evaluation focused on the Largest Connected Component, while community structure was assessed using Louvain modularity across 14 villages. All methods produced analyzable networks, achieving a 100% success rate and 97.17% average node coverage. Jaccard achieved the highest mean modularity (0.5074) and win rate (71.43%; 10 of 14 villages). A Friedman test confirmed differences among methods ($\chi^2 = 8.71$, $p = 0.013$). Holm-corrected Wilcoxon tests showed that Jaccard significantly outperformed Cosine and Pearson, whereas Cosine and Pearson did not differ significantly. These findings indicate that attribute-overlap-based edge construction provides most consistent representation of household welfare-profile proximity under the tested binary representation and automatic-thresholding scheme, while emphasizing that this advantage is context-dependent rather than universal.

Keywords: Data Desa Presisi; Edge Construction; Household Similarity Network; Jaccard Index; Louvain Modularity

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

Rural development requires data systems capable of portraying community welfare accurately, in detail, and in a form that can support decision-making. To date, much rural development data has been presented in tabular, aggregate, or spatial form, as reflected in the human development index and the village development index commonly used to capture the socio-economic conditions of communities [1], [2], [3]. Such presentations are valuable because they offer a general overview of a community's socio-economic situation. However, tabular and spatial approaches cannot fully explain patterns of proximity or similarity among households within a village's social structure. Yet in data-driven development, information about how closely households resemble one another can help reveal patterns of welfare clustering, social vulnerability, and the relational structure of rural communities.

Data Desa Presisi (Precision Village Data) has emerged as a village data-collection approach that emphasizes micro-level accuracy through the principles of by name, by address, and by coordinate. It integrates participatory census, drone-based spatial data, and a data-management system capable of generating information at the individual and household levels [4]. Beyond serving as an administrative database, Data Desa Presisi functions as a source of micro-level information that can be used to read community welfare conditions in far greater depth.

The need for micro-level data also resonates with critiques of top-down data approaches. Couldry argues that modern data relations often strip communities of their voice and agency [5], whereas the "big data from the bottom up" perspective casts local communities as key actors in the production and use of data [6]. Seen in this light, Data Desa Presisi becomes especially relevant because it treats the village as a site of data production that is more participatory, more contextual, and closer to the realities of community life.

Even so, the use of Data Desa Presisi has often centered on descriptive reporting, spatial visualization, and the mapping of welfare indicators. Such approaches are not enough to uncover the relational ties among households. Two households may live in different locations yet share very similar welfare profiles; conversely, households that are

geographically close are not necessarily alike in their socio-economic conditions. What is needed, then, is a computational approach that can transform household welfare data into a relational structure so that similarities among households can be analyzed systematically.

Social Network Analysis offers precisely such an approach. Within it, social entities are represented as nodes, while the relationships among them are represented as edges [7], [8]. Graph theory and social network analysis have been applied widely across both the computer and social sciences [9], including in efforts to support innovation and development in rural areas [10]. In this study, households are represented as nodes, but an edge is not taken to mean a direct social relationship such as communication, kinship, or personal interaction. Rather, an edge denotes a computational relationship based on the similarity of welfare profiles. The resulting network is therefore a welfare network built on attribute similarity.

It is important to distinguish between three related but different notions of a network. A social network in the strict sense records empirically observed social ties, such as communication, kinship, or interaction between actors. A similarity network, by contrast, connects entities whose attribute profiles resemble one another, regardless of whether those entities ever interact. The village welfare network built in this study is a similarity network of this second kind, and more specifically a welfare-proximity network: two households are linked not because they know or interact with one another, but because their welfare profiles are computationally close. This distinction matters for interpretation, because communities detected in such a network reflect shared welfare characteristics rather than observed social relationships.

A short example clarifies how two households become connected under each measure. After the five welfare dimensions of two households are discretized and one-hot encoded, Cosine Similarity links them when their binary profile vectors point in a similar direction, the Jaccard Index links them when the set of welfare categories they share is large relative to the categories they hold in total, and Pearson Correlation links them when their category values vary in a co-directional pattern. The same pair of households may therefore be judged similar by one measure and dissimilar by another, which is precisely why the choice of edge construction method is consequential.

The edge construction stage is a particularly critical part of building such a network. In similarity-based networks, edges are the foundation that determines network density, the legibility of communities, and the quality of any downstream analysis. If the edge construction method is poorly chosen, the resulting network may turn out too dense, too sparse, or unable to represent the welfare structure in any meaningful way. Selecting an appropriate similarity measure is thus a methodological decision of real consequence when building a village welfare network.

Several similarity measures can be used to construct edges, among them Cosine Similarity, the Jaccard Index, and Pearson Correlation. Cosine Similarity gauges the directional agreement between attribute vectors, the Jaccard Index gauges the proportion of shared attributes relative to their union, and Pearson Correlation gauges the alignment of variation patterns across attributes [11]. The choice of similarity measure is fundamental to graph-based analysis because it defines how proximity between entities is established [12]. Because these three methods rest on different measurement logics, they can yield different network structures even when applied to the very same data.

Prior work in Social Network Analysis has generally taken an already-formed network as its starting point for examining social structure through community detection, centrality measurement, analysis of connectivity patterns, and evaluation of community structure [7], [8], [13], [14]. In that tradition, the network is often treated as the primary object of analysis once nodes and edges have been defined. In similarity-based networks, however, the edge construction stage is itself the methodological step that shapes the network from the outset. Different similarity measures embody different measurement logics and can therefore produce different relational structures, edge densities, and degrees of community legibility [11]. Moreover, the modularity value used to evaluate community quality depends directly on the connection structure and edge weights of the network [15]. Edge construction should therefore be paired with an objective mechanism for selecting relationships for instance, automatic thresholding grounded in the principle of strong ties [16] so that the network becomes neither too dense nor too sparse. In rural network studies more generally, the analysis should also concentrate on the parts of the network that are genuinely connected, which makes the Largest Connected Component a relevant tool for preserving the integrity of the structure under study [17]. The resulting

edge weights cast the village welfare network as a weighted similarity network [18]. The intuition of connecting households that resemble one another is related to homophily, the tendency of entities with similar attributes to associate [19], [20]; here, however, homophily is invoked only as an analogy, because in its classical sense it describes the formation of empirically observed social ties among similar actors, whereas the ties in this study are constructed computationally from attribute similarity rather than observed directly. Comparing edge construction methods is thus an important first step before more complex network analyses such as Louvain community detection, centrality [21], assortativity [22], and network-based policy evaluation.

Although similarity-based graph construction is well established in the broader machine-learning literature for instance in the construction of k-nearest-neighbor and threshold (epsilon-neighborhood) graphs for spectral clustering [23], [24], [25] it has rarely been examined comparatively in the specific setting of household welfare data. Most applications of Social Network Analysis to rural development assume that the network already exists and pay little attention to how the choice of similarity measure shapes the very edges on which every subsequent analysis depends. This study addresses that gap by treating edge construction itself as the object of comparison.

Reading welfare similarity at the household level is useful for village policy in several concrete ways. Communities of households with similar welfare profiles can support the targeting of social assistance, cluster-based intervention design, the mapping of social vulnerability, and more equitable resource allocation. A welfare-similarity network makes these patterns visible in a relational form that tabular indices cannot, while remaining explicitly distinct from any claim about real social ties.

Geographic proximity is not used as a comparison baseline in this study because the research question concerns similarity in welfare attributes rather than spatial location; two households that are spatially adjacent may differ sharply in welfare, and a spatial baseline would answer a different question. Modularity is adopted as the principal evaluation indicator because it is a standard, method-agnostic measure of how well a weighted network partitions into internally cohesive communities, which makes it suitable for comparing edge construction methods on an equal footing. It should be read, however, as an indicator of structural quality rather than of policy validity: a higher

modularity means a more clearly partitioned network, not necessarily a more correct or more actionable welfare classification.

This study sets out to compare three edge construction methods for the Village Welfare Index network Cosine Similarity, the Jaccard Index, and Pearson Correlation. The comparison seeks to determine which method produces the most informative network structure, judged by modularity, LCC node coverage, success in forming the graph, the number of clusters, network density, and the consistency with which each method prevails across villages. Its main contribution is to provide a methodological basis for choosing an edge construction method suited to village welfare networks before more advanced analyses such as community detection, centrality, assortativity, or network-based policy evaluation are undertaken.

2. METHODS

2.1. Research Design

This study employs a quantitative-computational approach grounded in Social Network Analysis. Its central aim is to compare three edge construction methods within village welfare networks. Each village is analyzed as a network in its own right, so that the performance of each method can be judged against the structural characteristics of the network it produces.

The study is bounded to the stages of network construction, similarity computation, automatic threshold selection, extraction of the Largest Connected Component, community detection using the Louvain algorithm as an evaluation tool, and comparison of network-structure indicators. It therefore does not pursue downstream analyses such as centrality, assortativity, or policy auditing; instead, it concentrates on a single methodological question: which edge construction method is best suited to building a village welfare network based on attribute similarity?

The overall workflow of this study is summarized in Figure 1. The analysis proceeds in sequential stages, beginning with the household welfare data drawn from Data Desa Presisi. Each household is first represented as a vector of welfare attributes, which is then standardized through discretization and one-hot encoding so that every method

operates on a consistent data format. Pairwise similarity between households is then computed using three methods Cosine Similarity, the Jaccard Index, and Pearson Correlation and an edge is formed whenever the similarity value exceeds an automatically selected threshold. The resulting weighted graph is reduced to its Largest Connected Component, after which the Louvain algorithm detects communities and modularity is computed. Finally, the three methods are compared across all 14 villages using the evaluation indicators described in the following subsections.

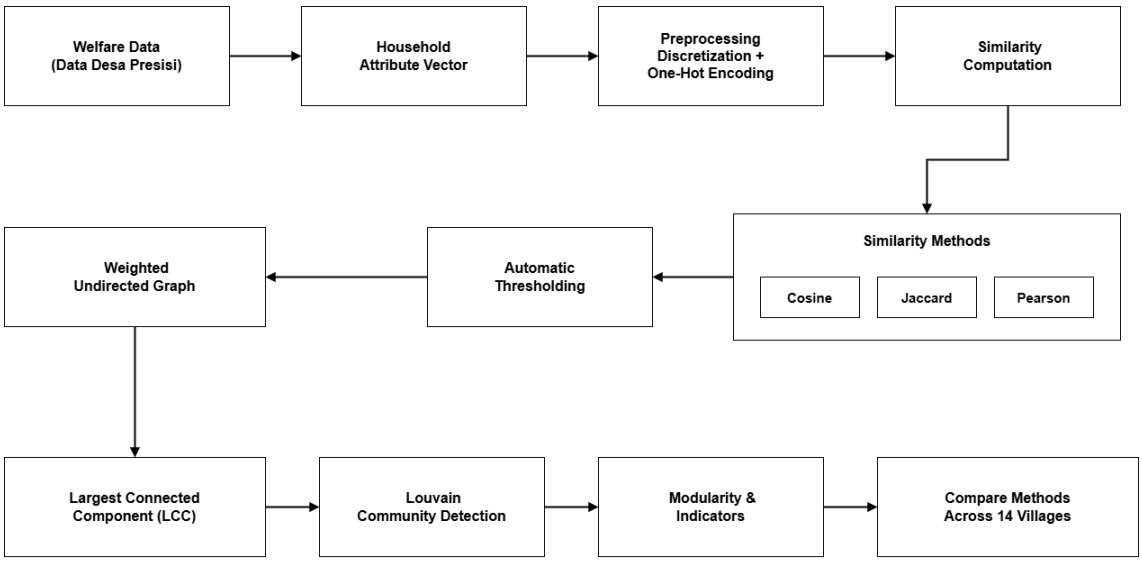

Figure 1. Research flowchart of the village welfare network analysis using three edge construction methods

2.2. Data Sources and Unit of Analysis

The data used in this study come from Data Desa Presisi, which records household welfare information that can be used to build a socio-economic profile of village communities. The unit of analysis is the household, or head of family. Households were chosen because community welfare in villages can be read more stably at the family level than at the individual level [4]. Each household is represented as a node in the network. The primary attributes used to form the relationships between nodes are the five dimensions of the Village Welfare Index, which together capture the basic aspects of household life the fulfillment of needs, education, social conditions, health, employment, environment, and infrastructure. Their composition follows the multidimensional framework for measuring human development [2], while the aggregate

Village Welfare Index score is computed using the geometric mean to avoid distortion from extreme values [26].

Table 1. Node-forming variables of the Village Welfare Index

No	Dimension	Function in the Model
1	Clothing, Food, and Shelter	Represents the fulfillment of basic household needs
2	Education and Culture	Represents educational capability and socio-cultural conditions
3	Social Life, Legal Protection, and Human Rights	Represents household social conditions and protection
4	Health, Employment, and Social Security	Represents health, employment, and socio-economic security conditions
5	Environment and Infrastructure	Represents environmental quality and access to basic infrastructure

This study draws on data from 14 villages: Bumi Ayu, Curup, Harapan Jaya, Lunas Jaya, Modong, Muara Dua, Muara Sungai, Pandan, Raja, Sedupi, Sukaraja (Tanah Abang), Tanah Abang Jaya, Tanah Abang Utara, and Tanjung Dalam.

2.3. Node Representation

Each household is represented as a vector of welfare attributes, as shown in Equation 1. Conceptually, the i-th household is expressed as:

$$X_i = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip}) \quad (1)$$

where X_i represents the welfare profile of the i-th household and p represents the number of attributes, or dimensions, used. It is this vector that serves as the basis for computing similarity between households.

Before similarity values are computed, the attribute data are standardized so that every household shares the same representational structure. The values across the five Village Welfare Index dimensions are processed through rounding, or discretization, at a given

level of precision in order to dampen the influence of very small numerical differences [27]. The data are then converted into a binary representation through one-hot encoding. This step ensures that every similarity method is applied to a consistent data format and can be compared fairly.

The standardization procedure is applied uniformly to every method. Each household is described by the five Village Welfare Index dimension scores (Table 1); these are dimension-level aggregate scores rather than their underlying sub-indicators, so that the comparison operates on the same five-dimensional welfare profile throughout. Each dimension score is first discretized by rounding to a fixed number of decimal places—rounding to the nearest integer (zero decimal places) in the main analysis—because household welfare scores often exhibit very fine, sub-decimal variation, yet such small differences do not necessarily represent substantive differences in a household's living conditions. Rounding therefore controls the granularity of the data so that micro-variation does not dominate the reading of the social structure, so that values differing only marginally fall into the same category. As a consequence, the analysis no longer rests on raw numerical differences but on the similarity of the welfare-category configurations that households share. Every distinct rounded value of a dimension is then treated as a category and expanded through one-hot encoding, so that the final feature vector of each household is a binary vector whose length equals the total number of distinct rounded categories observed across the five dimensions in that village; the exact number of binary features therefore varies by village according to the diversity of welfare values present, and is reported per village in Appendix B (Table B1). The sensitivity of the results to this rounding precision is examined in Section 3.9.

In graph terms, the model can be understood as a labeled graph: each node carries the welfare attributes of a household, while each edge carries the similarity weight between two nodes. The notion of a labeled graph is appropriate whenever the entities and relationships in a network carry attached attribute information [28].

2.4. Similarity-Based Edge Construction

Edges are constructed by computing the similarity between every pair of households within each village, as shown in Equation 2. If a village contains n households, the number of candidate pairs to be evaluated is:

$$n(n-1)/2 \quad (2)$$

Each pair of households receives a similarity value from one of the three methods under examination Cosine Similarity, the Jaccard Index, and Pearson Correlation. An edge is formed when that value meets the automatically selected threshold. In general terms, the rule for forming an edge is written as Equation 3:

$$(v_i, v_j) \in E \iff s_{ij} \geq t^* \quad (3)$$

where v_i and v_j are a pair of nodes, s_{ij} is the similarity value between node i and node j, and t^* is the automatically selected threshold obtained through the automatic thresholding rule (Algorithm 2). The edge weight is then set according to the similarity value, as shown in Equation 4:

$$w_{ij} = s_{ij} \quad (4)$$

Accordingly, the higher the similarity value, the stronger the welfare relationship between two households in the network.

2.5. Cosine Similarity

Cosine Similarity measures the directional similarity between two household vectors. It asks whether two households share a similar orientation in their welfare profiles, even when not all of their attributes are identical. The formula is given in Equation 5:

$$Sim_{cos}(P, Q) = (P \cdot Q) / (\|P\| \|Q\|) \quad (5)$$

A value approaching 1 indicates that two households have increasingly similar welfare-profile orientations [11], [29]. Cosine Similarity is used here because it captures broad similarities in pattern between households. It is well suited to cases where the goal is to identify households with comparable welfare orientations, even when their attributes are not identical in every respect.

2.6. Jaccard Index

The Jaccard Index measures the proportion of attributes that two households genuinely share. Unlike Cosine Similarity, which emphasizes vector direction, the Jaccard Index emphasizes the intersection, or overlap, of attributes. Its formula is given in Equation 6:

$$Sim_{jac}(P, Q) = |P \cap Q| / |P \cup Q| \quad (6)$$

In binary vector form, the same formula can be written as Equation 7:

$$Sim_{jac}(P, Q) = (P \cdot Q) / (\|P\|^2 + \|Q\|^2 - P \cdot Q) \quad (7)$$

The Jaccard Index rises as two households share more attributes relative to the total number of attributes they hold between them [11], [30]. As a result, it tends to be stricter in forming edges, since only relationships with substantial attribute overlap attain high values.

2.7. Pearson Correlation

Pearson Correlation measures the alignment of variation patterns between attribute vectors. It considers not only the magnitude of attribute similarity but also whether the patterns of change across attributes move in the same direction for two households. Its formula is given in Equation 8:

$$\rho(P, Q) = cov(P, Q) / (\sigma_P \sigma_Q) \quad (8)$$

where cov(P, Q) is the covariance between vectors P and Q, and σP and σQ are their respective standard deviations [11], [31]. Pearson Correlation is included as a comparative method because it reveals whether two households share a co-directional welfare pattern. It is useful for reading similarity in the structure of variation across attributes, rather than mere similarity in values or attribute overlap.

Table 2. Conceptual comparison of the three similarity methods

Method	Measurement Focus
Cosine Similarity	Directional similarity of vectors
Jaccard Index	Attribute intersection relative to attribute union
Pearson Correlation	Alignment of variation patterns

2.8. Automatic Thresholding and Weighted Graph Construction

Not every household pair with a similarity value is turned into an edge. A threshold is used to retain only relationships strong enough to keep the network from becoming either too dense or too sparse. In this study, the threshold is set through automatic thresholding based on the distribution of similarity values. This approach follows the principle of forming strong ties described by Cheng et al. [16], whereby only relationships above a certain level of closeness are preserved as strong ties in the network.

Technically, automatic thresholding proceeds by evaluating a range of candidate threshold values drawn from a fixed grid running from 0.1 to 0.9 in steps of 0.1. This grid is used because the similarity values are bounded in the interval [0, 1]; the extreme cutoffs 0 and 1 are excluded because they would retain, respectively, almost all or almost no pairs, and a step of 0.1 provides a resolution fine enough to separate near-complete from near-empty graphs while keeping the search inexpensive and fully reproducible. For each candidate threshold, the system counts the number of edges, or strong ties, that would form. The automatically selected threshold is the value whose retained-edge count is closest to the average retained-edge count across the whole grid [25]. This rule is adopted for a practical comparability reason rather than because it optimizes an external validity criterion: by targeting the grid-average retained-edge count it gives every similarity measure a network of comparable density, avoiding both near-complete graphs (threshold too low) and near-empty, disconnected graphs (threshold too high), so that the subsequent modularity comparison is not confounded by large density differences. In this way, the threshold is not set subjectively but is determined by the distribution of similarity values in each network.

Formally, as expressed in Equation 9, an edge is formed when the similarity between two households satisfies:

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

$$(v_i, v_j) \in E \Leftrightarrow s_{ij} \geq t^* \quad (9)$$

with the edge weight given by Equation 10:

$$w_{ij} = s_{ij} \quad (10)$$

where t^* is the automatically selected threshold derived from the distribution of similarity values, and w_{ij} is the weight of the relationship between node i and node j. The greater the similarity, then, the stronger the weighted tie between two households in the graph. The outcome of this stage is an undirected weighted graph that represents the structure of household welfare similarity more objectively.

2.9. Largest Connected Component

Once the graph has been formed, the Largest Connected Component (LCC) is extracted. The LCC is the largest component of the network in which every node remains reachable from the others through some path. Extracting it focuses the analysis on the principal part of the network the portion with the strongest relational integrity. This approach follows Montes et al. [17], who use connected network components to read social structure in a more functional way. Here, the LCC ensures that the evaluation of communities is not unduly influenced by isolated nodes or small components that do not represent the main structure of the village network.

This step matters because a thresholded network can contain several small components. If all of them were kept in the evaluation, the modularity value could be distorted by parts of the network that do not represent its main structure. Louvain community detection and modularity evaluation are therefore performed on the LCC. Its node coverage is computed using the formula in Equation 11:

$$\text{Coverage} = (N_{LCC} / N_{village}) \times 100\% \quad (11)$$

where N_LCC is the number of nodes in the largest connected component and N_village is the total number of household nodes in the village. This coverage value confirms that using the LCC does not remove too large a share of the data from the analysis.

2.10. Community Detection and Modularity Evaluation

Once the LCC network is in place, its community structure is analyzed using the Louvain algorithm. In this study, Louvain is not the main focus but rather a tool for obtaining the modularity value, which gauges how strongly the network divides into well-defined communities. Modularity captures the extent to which nodes within the same community are more strongly connected than would be expected by chance [15]. It is expressed in Equation 12:

$$Q = (1/2m) \Sigma_{ij} [A_{ij} - k_i k_j / 2m] \delta(c_i, c_j) \quad (12)$$

where A_{ij} is the edge weight between nodes i and j, k_i and k_j are the total edge weights at each node, m is the total edge weight in the network, and $\delta(c_i, c_j)$ equals 1 when nodes i and j belong to the same community. The Louvain algorithm is used because it efficiently optimizes modularity in large networks and has been widely adopted in comparisons of community-detection algorithms [14].

2.11. Indicators for Method Comparison

The three methods are compared using several key indicators: modularity per village, mean modularity, mean node coverage, success rate, win rate, number of clusters, number of edges, number of LCC nodes, and density. Together, these indicators read each method's performance in terms of community-structure quality, network coverage, and the consistency of results across villages.

Table 3. Indicators for evaluating edge construction methods

Indicator	Function
Modularity per village	Assesses the quality of community structure in each village
Mean modularity	Measures a method's average performance across all villages
Mean coverage node	Assesses the share of nodes that enter the LCC

Indicator	Function
Success rate	Measures a method's success in forming an analyzable graph
Win rate	Measures how often a method is the best performer
Number of clusters	Reads the number of communities produced by Louvain
Number of edges	Measures the number of similarity relationships formed
Density	Measures the density of the network
LCC nodes	Measures the number of nodes within the largest connected component

2.12. Algorithm, Implementation, Reproducibility, and Data Ethics

To make the procedure fully reproducible, the complete pipeline and the automatic threshold-selection rule are stated as pseudocode in Algorithm 1 and Algorithm 2. The pipeline is applied independently to every village, and villages with fewer than five households are excluded because they cannot form a meaningful network.

Algorithm 1. Village welfare similarity-network construction and community detection

```
Algorithm 1: Village Welfare Similarity-Network Construction and Community Detection
-------------------------------------------------------------------------------
Input:
  D_v   : household (kepala keluarga) records for village v
  F     : welfare indicator columns {f_A, f_B, f_C, f_D, f_E} (5 IKD dimensions)
  M     : similarity method, M in {Cosine, Jaccard, Pearson}
  r     : one-hot rounding precision, r in {0, 1, 2} decimal places
  T     : threshold grid, T = {0.1, 0.2, ..., 0.9}
  seed  : Louvain random seed (default = 42)

Output:
  G_LCC : largest connected component of the similarity graph
  C     : Louvain community partition of G_LCC
  Q     : modularity of C on G_LCC
  t*    : selected similarity threshold

 1:  D_v <- FilterHeadsOfHousehold(D_v)                  // subjek contains "kepala keluarga"
 2:  D_v <- DropDuplicates(D_v, key = family_id)
 3:  if |D_v| < 5 then
 4:      return NULL                                     // insufficient households to form a
network
 5:  end if
 6:  X <- OneHotEncode( Round(D_v[F], r) )               // binary feature matrix, one column-
set per rounded indicator
 7:  E_cand <- empty list                                // candidate edges (all unordered
pairs)
 8:  for i <- 1 to |D_v| - 1 do
 9:      for j <- i + 1 to |D_v| do
10:         w_ij <- Similarity_M( X[i], X[j] )           // Cosine, Jaccard, or Pearson
(Algorithm 1a-1c)
11:         append (i, j, w_ij) to E_cand
12:     end for
13: end for
```

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

```
14: t* <- SelectThreshold( {w : (i,j,w) in E_cand}, T )     // Algorithm 2
15: G <- Graph(nodes = D_v)                                 // one node per household, no edges
yet
16: for each (i, j, w_ij) in E_cand do
17:     if w_ij >= t* then
18:         AddEdge(G, i, j, weight = w_ij)
19:     end if
20: end for
21: if |Edges(G)| = 0 then
22:     return NULL                                         // no edges survive thresholding
23: end if
24: G_LCC <- LargestConnectedComponent(G)
25: C_raw <- Louvain(G_LCC, weight = "weight", random_state = seed)
26: C <- ReorderClusterLabels(C_raw, key = mean(basis_indicator) per cluster)  // ascending
welfare score; cosmetic only
27: Q <- Modularity(C, G_LCC, weight = "weight")
28: return (G_LCC, C, Q, t*)
```

```
Algorithm 1a: Cosine Similarity
-------------------------------
Input: binary feature vectors x_i, x_j
1: denom <- ||x_i|| * ||x_j||
2: if denom <= 1e-12 then return 0
3: return clip( (x_i . x_j) / denom , 0, 1 )
```

```
Algorithm 1b: Jaccard Index
---------------------------
Input: binary feature vectors x_i, x_j
1: A <- { k : x_i[k] > 0 },  B <- { k : x_j[k] > 0 }
2: if |A union B| = 0 then return 0
3: return |A intersect B| / |A union B|
```

```
Algorithm 1c: Pearson Correlation
---------------------------------
Input: binary feature vectors x_i, x_j
1: if std(x_i) <= 1e-12 or std(x_j) <= 1e-12 then return 0
2: rho <- corrcoef(x_i, x_j)
3: if rho is not finite then return 0
4: return rho   // negative values are not rescaled/discarded here; they are filtered out
                // downstream by thresholding (step 17), since t* in this study is always in
[0.1, 0.9]
```

Algorithm 2. Automatic similarity-threshold selection

```
Algorithm 2: Automatic Similarity-Threshold Selection ("Auto Distribution" mode)
--------------------------------------------------------------------------------
Input:
  S : multiset of all candidate pairwise similarity values for one village x one method
  T : threshold grid, T = {0.1, 0.2, ..., 0.9}  (step = 0.1)

Output:
  t* : selected threshold

 1: if S = empty then
 2:     return 0.4                                       // fallback default
 3: end if
 4: for each t in T do
 5:     count[t] <- |{ w in S : w >= t }|                // number of edges retained if
threshold = t
 6: end for
 7: total <- sum( count[t] for t in T )
 8: target <- total / |T|                                // average retained-edge count across
the whole grid
 9: for each t in T do
10:     score[t] <- ( |count[t] - target|, |t - 0.5| )    // lexicographic key: (a) closeness
to the average
```

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

```
11:                                                    //   retained-edge count, (b) tie-
break toward 0.5
12: end for
13: t* <- argmin over t in T of score[t]               // smallest score, lexicographic
order
14: return t*

Rationale: the grid search favors a threshold whose retained-edge count is closest to what an
"average"
threshold in the grid would retain, avoiding both near-complete graphs (t too low) and near-
empty /
disconnected graphs (t too high), without requiring a manually tuned, village-specific cutoff.
```

Several implementation details follow directly from this construction. Candidate edges are generated only for unordered household pairs i < j, so self-loops cannot arise, and each pair is added to the graph exactly once through an idempotent undirected-edge operation, so duplicate edges cannot arise. Modularity is computed on the weighted graph, using the similarity values (Cosine, Jaccard, or Pearson) as edge weights rather than treating the graph as unweighted. Negative Pearson correlations are neither discarded nor rescaled; they are retained as computed and then filtered out naturally at the thresholding stage, because the selected threshold in every village lies in the range 0.1-0.9, so virtually all negatively correlated pairs fall below it and form no edge. A fixed random seed (random_state = 42) is used for the main analysis to guarantee reproducibility, and the sensitivity of modularity to this seed is examined explicitly in Section 3.11.

Because household welfare records are sensitive micro-level data collected by name, by address, and by coordinate, they are handled with corresponding care. In this study each household is referenced only through an internal family identifier rather than a national identity number or personal name, and any welfare micro-data released as supporting material should be anonymized further (for example by hashing the family identifier) before public deposit. The analysis reports only aggregate, network-level results, and no individually identifying information is presented.

Table 4. Software and library versions used for reproducibility

Library / Tool	Version
Streamlit	1.54.0
pandas	2.3.3
NumPy	2.4.2
SciPy	1.17.0

Library / Tool	Version
scikit-learn	1.8.0
NetworkX	3.3
python-louvain (community)	0.16
Plotly	6.5.2
Matplotlib	3.9.2
openpyxl	3.1.5
pyarrow	23.0.0
WordCloud	1.9.6
Sastrawi	1.0.1

Table 5. Network summary per village households before and after Largest Connected Component extraction, node coverage, edge density, and number of edges (the per-village LCC node counts, density, and edge counts are reported once because they are matched across the three methods; as explained in Section 3.4, this reflects equal edge counts and density at the automatically selected thresholds, not necessarily identical edge sets)

Table 5. Network summary per village

Village	Households (before LCC)	Households in LCC	Node coverage	Density	Edges
Bumi Ayu	193	179	92.7%	0.0358	570
Curup	502	495	98.6%	0.0223	2,729
Harapan Jaya	684	681	99.6%	0.0373	8,633
Lunas Jaya	318	315	99.1%	0.0422	2,087
Modong	421	418	99.3%	0.0276	2,408
Muara Dua	230	210	91.3%	0.0288	631
Muara Sungai	210	190	90.5%	0.0315	565
Pandan	853	842	98.7%	0.0313	11,092
Raja	775	775	100.0%	0.0481	14,430
Sedupi	541	534	98.7%	0.0271	3,856
Sukaraja (Tanah Abang)	315	314	99.7%	0.0355	1,746
Tanah Abang Jaya	876	873	99.7%	0.0525	19,990
Tanah Abang Utara	759	756	99.6%	0.0298	8,518
Tanjung Dalam	229	213	93.0%	0.0413	932

3. RESULTS AND DISCUSSION

3.1. Aggregate Performance Summary Across Methods

The analysis shows that all three edge construction methods are able to produce analyzable networks. This is evident from the success rate of 100% for Cosine Similarity, the Jaccard Index, and Pearson Correlation alike. In other words, every method succeeded in forming a graph for each of the 14 villages analyzed.

Table 6. Summary of aggregate indicators for edge construction methods

Method	Mean Modularity	Mean Coverage Node	Success Rate	Win Rate
Cosine Similarity	0.5031	97.17%	100%	14.29%
Jaccard Index	0.5074	97.17%	100%	71.43%
Pearson Correlation	0.5028	97.17%	100%	14.29%

Table 6 shows that the main difference between the methods does not lie in their success at forming graphs, since all three share the same success rate. Nor does it appear in LCC node coverage, as all three yield the same average node coverage of 97.17%. This indicates that the automatic thresholding and LCC extraction mechanisms keep the network stable and prevent too many nodes from dropping out of the main analysis. The real difference emerges instead in mean modularity and win rate. The Jaccard Index produces the highest mean modularity at 0.5074, followed by Cosine Similarity at 0.5031 and Pearson Correlation at 0.5028. The Jaccard Index also records the highest win rate at 71.43%, meaning it was the best-performing method in 10 of the 14 villages. Cosine Similarity and Pearson Correlation each post a win rate of 14.29%, prevailing in 2 villages apiece. Figure 2 shows that the Jaccard Index is the method most consistent in producing the highest modularity. Although the average modularity gap among the three methods is not large, the Jaccard Index's repeated wins across most villages suggest that it is more stable in forming welfare community structures.

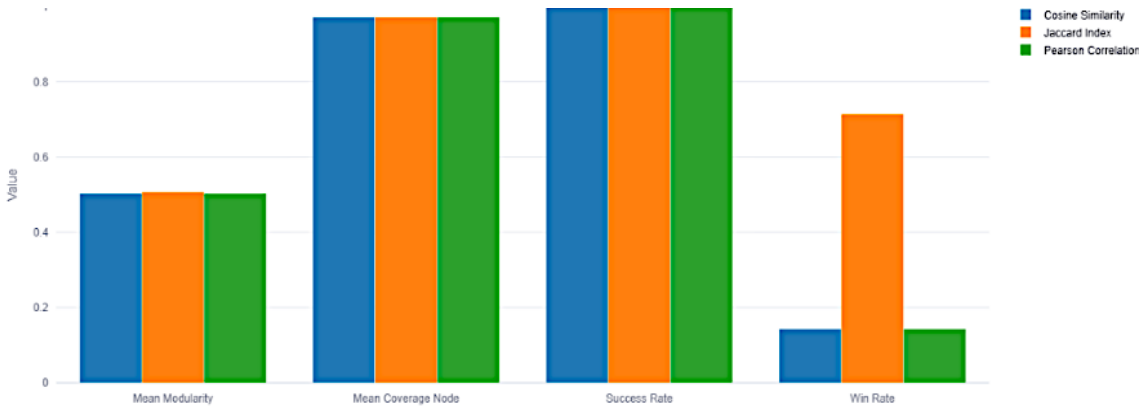

Figure 2. Comparison of aggregate indicators across edge construction methods

3.2. Comparison of Modularity Values Across Villages

Comparing modularity values across villages shows that the three methods produce relatively close value patterns. There are differences, however, in which method yields the highest value in each village.

Table 7. Modularity values of the village welfare networks under the three edge construction methods

No	Village	Cosine Similarity	Jaccard Index	Pearson Correlation
1	Bumi Ayu	0.6307	0.6303	0.6289
2	Curup	0.5278	0.5305	0.5356
3	Harapan Jaya	0.4549	0.4585	0.4500
4	Lunas Jaya	0.4532	0.4592	0.4542
5	Modong	0.5090	0.5138	0.5134
6	Muara Dua	0.5787	0.5783	0.5788
7	Muara Sungai	0.5435	0.5491	0.5423
8	Pandan	0.4818	0.4839	0.4826
9	Raja	0.4302	0.4421	0.4293
10	Sedupi	0.5078	0.5096	0.5082
11	Sukaraja (Tanah Abang)	0.5220	0.5282	0.5226
12	Tanah Abang Jaya	0.4115	0.4066	0.3999
13	Tanah Abang Utara	0.4640	0.4752	0.4673

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

No	Village	Cosine Similarity	Jaccard Index	Pearson Correlation
14	Tanjung Dalam	0.5279	0.5377	0.5266

Table 7 shows that the Jaccard Index is the method most frequently producing the highest modularity. It prevails in 10 villages—Harapan Jaya, Lunas Jaya, Modong, Muara Sungai, Pandan, Raja, Sedupi, Sukaraja (Tanah Abang), Tanah Abang Utara, and Tanjung Dalam. Cosine Similarity leads in 2 villages, Bumi Ayu and Tanah Abang Jaya, while Pearson Correlation also leads in 2, Curup and Muara Dua. In Bumi Ayu, the highest modularity comes from Cosine Similarity at 0.6307, slightly above the Jaccard Index at 0.6303 and Pearson Correlation at 0.6289. The very small gap suggests that the welfare structure of this village can be read fairly consistently by all three methods.

In Curup, Pearson Correlation yields the highest value at 0.5356, compared with 0.5305 for the Jaccard Index and 0.5278 for Cosine Similarity. This suggests that the welfare structure of Curup is read more strongly through the alignment of variation patterns across attributes than through vector direction or attribute intersection. In most other villages, by contrast, the Jaccard Index produces the highest modularity. This finding suggests that village welfare networks are generally better built on genuinely overlapping attributes. Put differently, two households are more appropriately treated as close in welfare when they share the same substantive combination of attributes.

Figure 3. Heatmap of modularity values by village and method

Figure 3 shows that the three methods follow closely matching modularity patterns across nearly every village. When a village has high modularity, all three methods tend to produce high values; when a village has low modularity, all three decline together. This pattern indicates that the characteristics of each village's data also shape the strength of the community structure that forms. Even so, the Jaccard Index tends to sit slightly above the other two. Its edge is not always large in numerical terms, but it is consistent across most villages. In network analysis, even small consistency at the edge construction stage matters, because it can influence downstream analyses such as community detection, centrality, assortativity, and structure-based policy evaluation.

3.3. Number of Clusters Produced by Each Method

The number of clusters produced by Louvain is used to observe how each method shapes the community structure of the village welfare network. In general, the number of clusters formed falls within a range of 7 to 11.

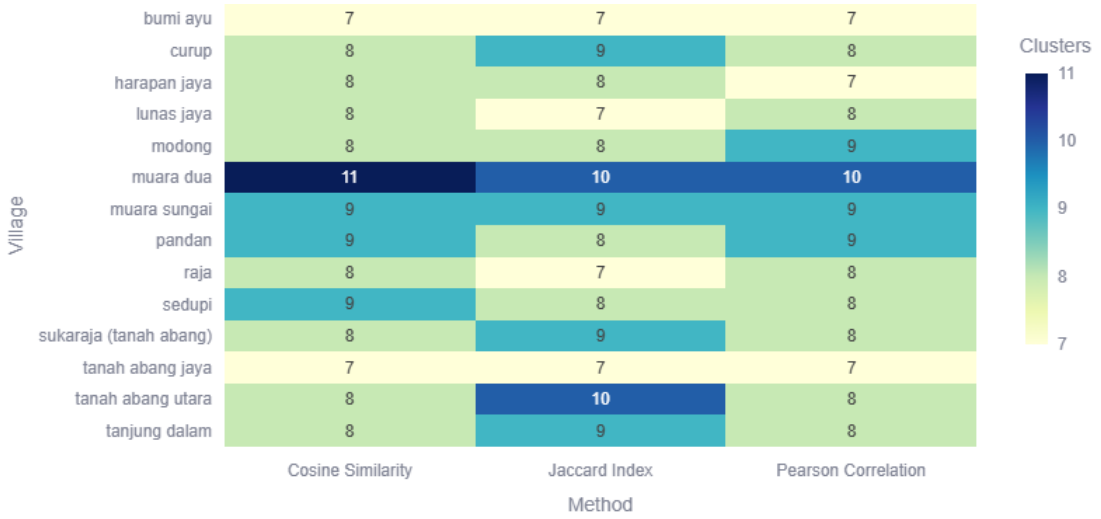

Figure 4. Number of clusters by village and method

Figure 4 shows that the number of clusters produced by the three methods is relatively stable. Cosine Similarity averages about 8.29 clusters, the Jaccard Index likewise averages about 8.29, and Pearson Correlation averages about 8.14. This indicates that the three methods do not differ dramatically in the number of communities they generate. There is, however, some variation in particular villages. In Muara Dua, for instance, Cosine Similarity produces 11 clusters, while the Jaccard Index and Pearson Correlation produce 10. In Tanah Abang Utara, the Jaccard Index produces 10 clusters, whereas Cosine

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Similarity and Pearson Correlation produce 8. This variation shows that the edge construction method can influence Louvain's sensitivity in detecting the separation of communities. Too few clusters can indicate that the network structure is overly merged, while too many can signal fragmentation. In this study, the relatively stable cluster counts indicate that automatic thresholding succeeded in keeping the network structure within an analyzable range.

3.4. Density, Number of Edges, and Number of LCC Nodes

Density is used to read how dense the network is. The results show that network density falls within a range of 0.0223 to 0.0525, indicating that village welfare networks are sparse. This matters because a network that is too dense makes communities hard to read, while one that is too loose can leave too many isolated nodes.

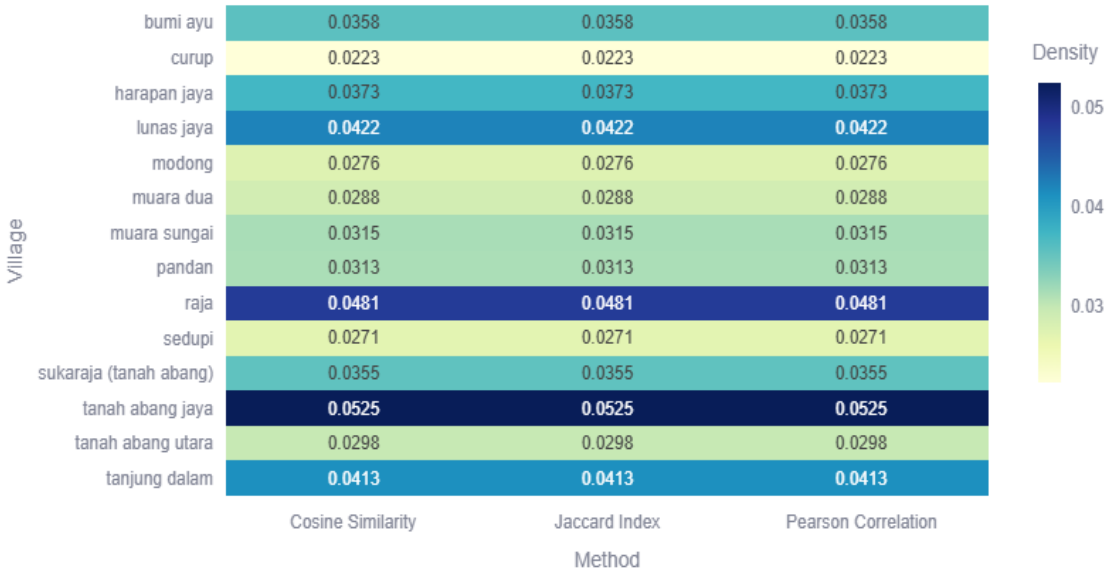

Figure 5. Density by village and method

Figure 5 shows that density is relatively similar across the three methods within each village. This indicates that automatic thresholding succeeded in keeping network density at comparable levels. As a result, the modularity comparison between methods becomes fairer, since differences in the results are not driven by extreme differences in density.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

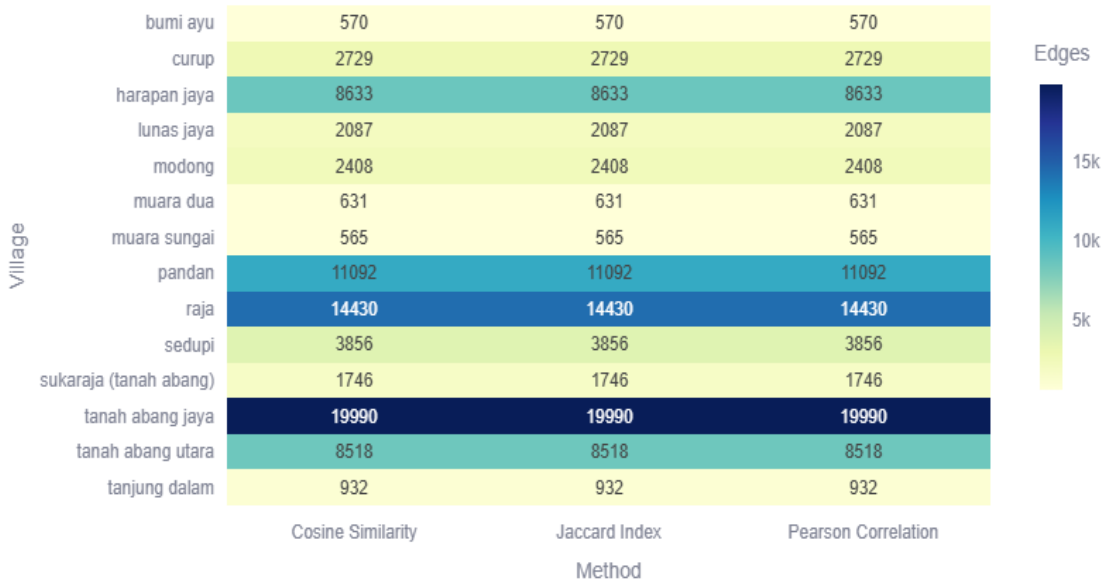

Figure 6. Number of edges by village and method

Figure 6 shows that the number of edges differs across villages but is relatively similar across methods within the same village. The lowest edge count occurs in Muara Sungai with 565 edges, while the highest occurs in Tanah Abang Jaya with 19,990 edges. This difference is driven mainly by the number of households in each village and by the similarity structure that emerges after thresholding. The exact per-village density and edge counts underlying Figures 5 and 6 are reported in Table 5. A natural question is why Table 5 reports the same number of edges, the same density, and the same LCC node counts for all three methods even though the similarity measures differ. What is matched across methods is the edge count and the resulting density, not necessarily the identity of the connected pairs: the automatic thresholding rule selects, for each measure, a cutoff that retains a comparable number of strong ties, so that the three village networks are compared at the same number of edges and the same density. The modularity differences analyzed below therefore do not stem from how many edges each method forms, but from how each measure weights the retained ties—and, where the retained pairs differ, from which ties are formed—so that weighted Louvain optimization yields different community partitions at matched density.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

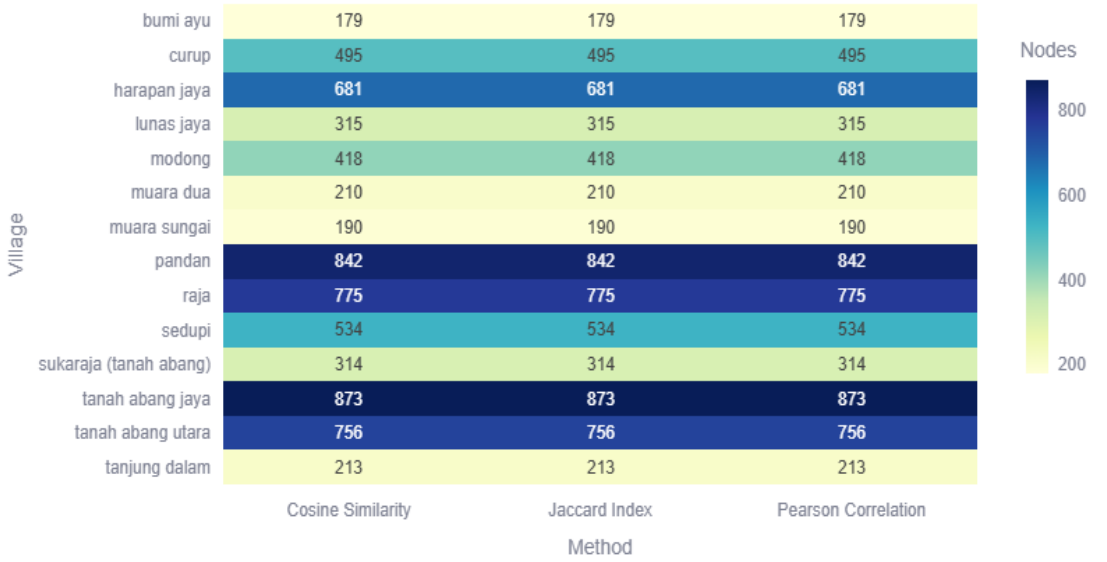

Figure 7. Number of LCC nodes by village and method

Figure 7 shows that the number of nodes in the LCC also varies across villages. The lowest value is in Bumi Ayu with 179 LCC nodes, while the highest is in Tanah Abang Jaya with 873. Overall, LCC node coverage is fairly high, averaging 97.17%. This means that most households remain part of the main network under analysis. These results reinforce that using the LCC does not discard the bulk of the data but simply focuses the analysis on the most relevant component of the network. With high average coverage, the community structure evaluated through modularity can still be regarded as representative of most households in a village.

3.5. Performance Characteristics of Cosine Similarity

Cosine Similarity performs competitively across all villages. The modularity values it produces are almost always close to those of the Jaccard Index and Pearson Correlation, which indicates that the direction of household welfare profiles is fairly consistent in shaping the network structure. Its main strength lies in reading the similarity of orientation between vectors: if two households have attribute patterns that move in the same direction, they are treated as similar even when their attributes are not all identical. Cosine Similarity is therefore well suited to capturing broad similarities in welfare profiles.

In this study, however, Cosine Similarity is not the method that prevails most often. It produces the highest modularity in only 2 of the 14 villages. This may be because village

welfare networks call for a stricter definition of relationships. Edges formed solely on the basis of profile direction can link households that appear broadly similar yet do not necessarily share strong attribute overlap. As a result, the communities that emerge can be a little less sharply defined than those built with the Jaccard Index.

3.6. Performance Characteristics of the Jaccard Index

The Jaccard Index is the most dominant method in this study. Most villages attain their highest modularity when edges are formed with it, which indicates that an intersection-based approach is better suited to representing welfare relationships between households. Its advantage can be traced to the character of household welfare data, which has been represented in categorical or binary form. In that format, the most meaningful similarity is similarity in attributes that are genuinely shared. The Jaccard Index directly computes the proportion of shared attributes relative to all the attributes appearing in two households.

It is therefore more selective in deciding whether two households deserve to be connected in the network. In village welfare networks, this selectivity matters. If edges form too easily, the network can become too dense and communities hard to distinguish. Conversely, when edges form only between households with strong attribute overlap, the resulting communities tend to be clearer. The results show that the Jaccard Index has the highest mean modularity at 0.5074 and the highest win rate at 71.43%. It can thus be regarded as the most consistent edge construction method for village welfare networks based on categorical or binary attributes.

3.7. Performance Characteristics of Pearson Correlation

Pearson Correlation performs relatively steadily, though not as dominantly as the Jaccard Index. In some villages its values come very close to those of the other two methods, and it even attains the highest value in Curup and Muara Dua. This indicates that in certain villages the welfare structure is shaped more by the alignment of patterns across attributes than by direct attribute similarity. Pearson Correlation excels at reading linear relationships between vectors: if two households have attribute variation patterns that move in the same direction, they receive a high correlation value. This approach can become less strict, however, when the data are categorical or the product of binary transformation. In such cases, a correlation in pattern does not necessarily mean that

two households truly share the same welfare attributes. For this reason, Pearson Correlation is better positioned as a comparative method. It is useful for checking whether community structure still emerges when edges are formed on the basis of aligned variation patterns. The results show that Pearson Correlation still forms good networks, but does not always produce the highest modularity.

3.8. Statistical Significance and Effect Sizes

Because the modularity differences among the three methods are numerically small, their statistical significance was tested directly rather than inferred from mean values alone. Treating the 14 villages as paired observations, a Friedman test across the three methods was significant (chi-square = 8.71, p = 0.013), indicating that at least one method differs systematically from the others. Pairwise Wilcoxon signed-rank tests with Holm correction (Table 8) show that the Jaccard Index significantly outperforms both Cosine Similarity (Holm-adjusted p = 0.013) and Pearson Correlation (Holm-adjusted p = 0.012), with large rank-biserial effect sizes (r = 0.79 and r = 0.83, respectively). Cosine Similarity and Pearson Correlation do not differ significantly (p = 0.90). The distribution of per-village modularity for each method is shown in Figure 8; the boxes overlap substantially, which is consistent with the small absolute differences, but the paired tests confirm that Jaccard's advantage is systematic rather than incidental.

Table 8. Pairwise Wilcoxon signed-rank comparisons of modularity across the 14 villages, with Holm-adjusted p-values and rank-biserial effect sizes

Comparison	N	Median d-mod.	W	p (raw)	r	p (Holm)	Sig.
Cosine Similarity vs Jaccard Index	14	-0.0042	11.0	0.0067	-0.79	0.0134	Yes
Cosine Similarity vs Pearson Correlation	14	-0.0002	50.0	0.9032	0.05	0.9032	No
Jaccard Index vs Pearson Correlation	14	0.0053	9.0	0.0040	0.83	0.0121	Yes

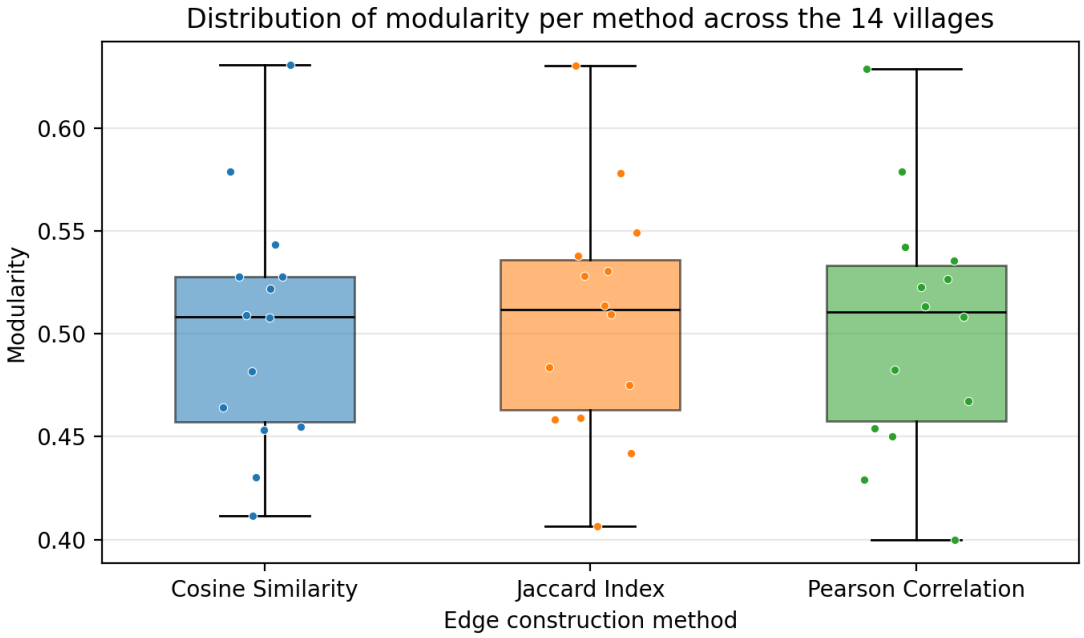

Figure 8. Distribution of modularity per method across the 14 villages

3.9. Robustness to Discretization Level

To test whether Jaccard's advantage depends on the chosen rounding precision, the analysis was repeated with the welfare dimensions discretized to integer (0 decimal), one-decimal, and two-decimal precision. As Table 9 and Figure 9 show, the Jaccard Index retains the highest mean modularity at every discretization level (for example 0.5704 at one decimal versus 0.5672 for Cosine and 0.5671 for Pearson). Coarser discretization lowers modularity for all three methods roughly in parallel, with integer rounding producing the lowest values (around 0.50). The fact that modularity responds systematically to discretization, rather than remaining fixed, indicates that the detected community structure reflects genuine variation in welfare profiles and is not merely an artifact of the one-hot encoding; at the same time, the ranking of the three methods is stable across all levels tested.

Table 9. Mean modularity per method under three discretization levels

Discretization level	Cosine Similarity	Jaccard Index	Pearson Correlation
1 decimal place(s)	0.5672	0.5704	0.5671
2 decimal place(s)	0.5666	0.5697	0.5662
Integer (0 decimal)	0.5031	0.5074	0.5028

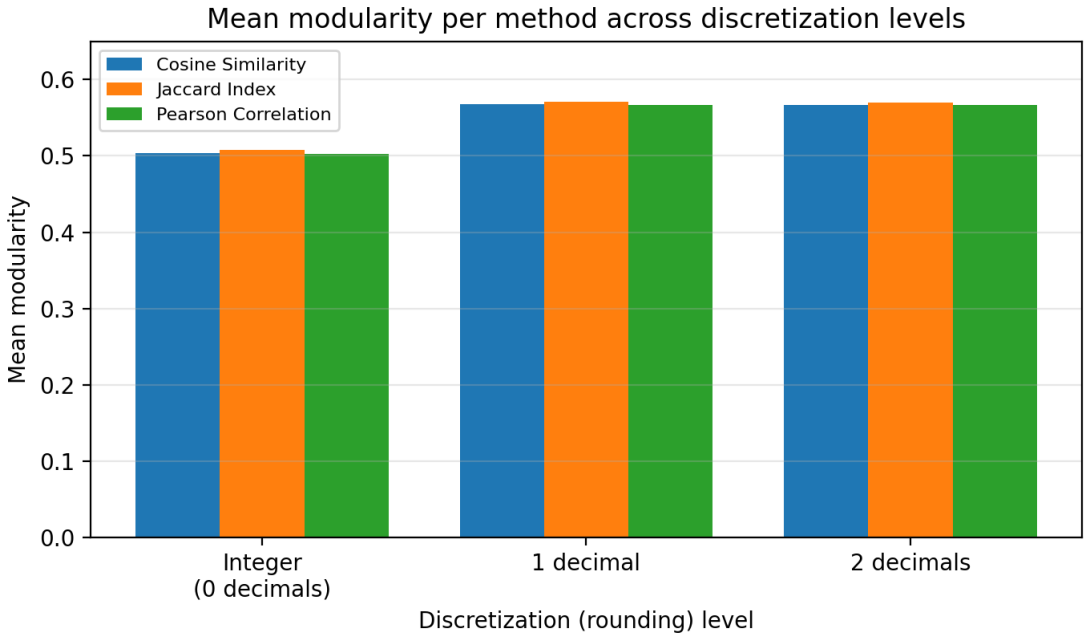

Figure 9. Mean modularity per method across discretization levels

3.10. Robustness to Graph-Construction Scheme

The comparison was further tested against two alternative graph-construction schemes that do not rely on the automatic similarity threshold: a k-nearest-neighbor graph (k = 8) [24] and a fixed-density graph matched to the density of the thresholded network [23]. Table 10 and Figure 10 report the mean modularity of each method under each scheme. Under the automatic-threshold scheme used in the main analysis, the Jaccard Index leads (0.5074, versus 0.5031 for Cosine and 0.5028 for Pearson). Under the fixed-density scheme the three methods are very close, with the Jaccard Index marginally ahead (mean modularity 0.4951, versus 0.4943 for Pearson Correlation and 0.4904 for Cosine Similarity). Under the k-nearest-neighbor scheme, however, the picture changes markedly: Pearson Correlation attains the highest mean modularity (0.5181) and wins in 10 of the 14 villages, while Jaccard leads in only four. This result is important for interpretation: it shows that the superiority of the Jaccard Index is specific to the similarity-thresholding construction and to the binary representation, and should not be generalized into a claim that Jaccard is universally the best similarity measure for welfare networks.

Table 10. Mean modularity per method under three graph-construction schemes (means across the 14 villages)

Graph-construction scheme	Cosine Similarity	Jaccard Index	Pearson Correlation
Automatic threshold (main)	0.5031	0.5074	0.5028
k-nearest-neighbor (k = 8)	0.4881	0.4957	0.5181
Fixed density	0.4904	0.4951	0.4943

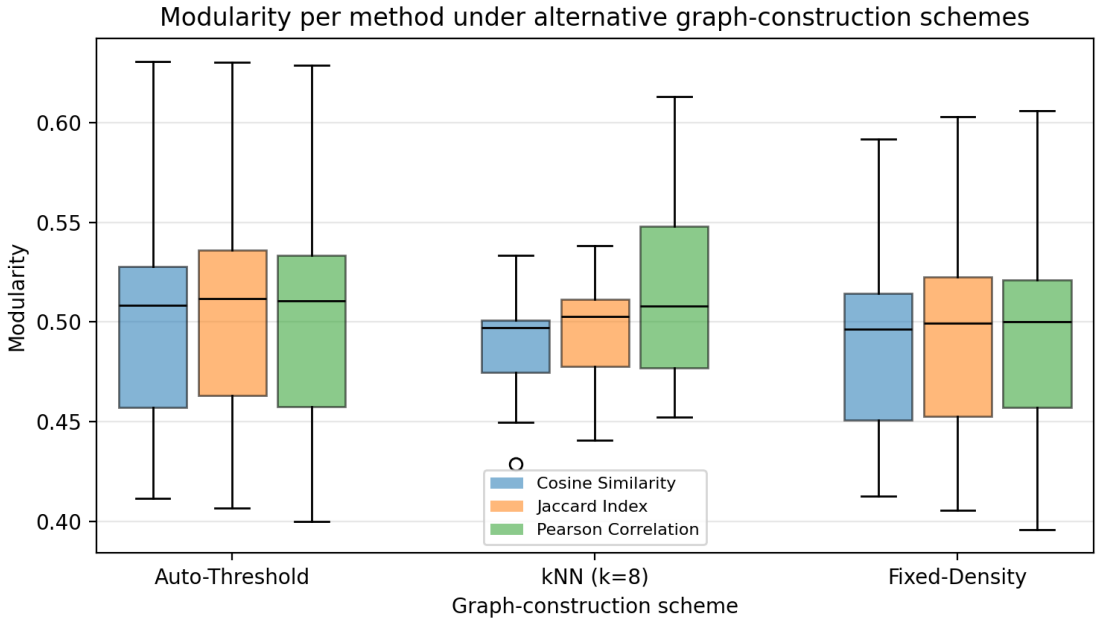

Figure 10. Modularity per method under alternative graph-construction schemes

3.11. Louvain Stability, Threshold Values, and the Meaning of Modularity

Because the Louvain algorithm is stochastic and can produce badly connected communities [32], its modularity output can vary between runs. To quantify this, community detection was repeated 30 times per village and method with different random initializations. The resulting variability is small: the mean coefficient of variation of modularity is 1.39% for Cosine Similarity, 1.42% for Jaccard Index, and 1.41% for Pearson Correlation (Table 11), and the mean modularity over repeats (0.5014, 0.5044, and 0.5018 respectively) preserves the same ranking as the fixed-seed analysis. The fixed random seed (random_state = 42) used in the main analysis therefore does not bias the comparison.

Table 11. Stability of Louvain modularity over 30 repeated runs per village and method

Method	Mean modularity (30 runs)	Mean SD	Mean CV
Cosine Similarity	0.5014	0.0068	1.39%
Jaccard Index	0.5044	0.0069	1.42%
Pearson Correlation	0.5018	0.0069	1.41%

The automatic threshold-selection rule (Algorithm 2) produced highly stable cutoffs: across all 14 villages it selected a similarity threshold of 0.3 for Cosine Similarity, 0.2 for the Jaccard Index, and 0.3 for Pearson Correlation (Table 12). The uniformity of these values indicates that the rule is not fitting noise in individual villages but is responding to a consistent property of the similarity distributions, and it confirms that the density differences between methods reported in Section 3.4 stem from the measures themselves rather than from divergent thresholds.

Table 12. Automatically selected similarity threshold per village and method

Village	Cosine Similarity	Jaccard Index	Pearson Correlation
Bumi Ayu	0.3	0.2	0.3
Curup	0.3	0.2	0.3
Harapan Jaya	0.3	0.2	0.3
Lunas Jaya	0.3	0.2	0.3
Modong	0.3	0.2	0.3
Muara Dua	0.3	0.2	0.3
Muara Sungai	0.3	0.2	0.3
Pandan	0.3	0.2	0.3
Raja	0.3	0.2	0.3
Sedupi	0.3	0.2	0.3
Sukaraja (Tanah Abang)	0.3	0.2	0.3
Tanah Abang Jaya	0.3	0.2	0.3
Tanah Abang Utara	0.3	0.2	0.3
Tanjung Dalam	0.3	0.2	0.3

Finally, the practical meaning of the modularity values deserves comment. Values around 0.50, combined with 7 to 11 detected communities per village, indicate a moderately

strong and clearly identifiable partition: households fall into welfare groups that are appreciably more cohesive internally than would be expected by chance, without the network fragmenting into many tiny components. In the welfare context this suggests the presence of several distinct welfare strata within a village rather than a single homogeneous population. This structural reading is deliberately cautious: modularity describes how cleanly the network divides and is itself subject to a resolution limit that can obscure very small communities [33], so confirming that these strata correspond to recognized welfare or poverty categories would require external validation with village stakeholders, which lies beyond the present scope.

3.12. Methodological Implications

The results of this study show that edge construction should not be treated as a purely technical step. In similarity-based networks, edges are the foundation of the entire analysis. Different methods can produce different community structures even when the underlying data are the same. Automatic thresholding offers a methodological advantage because the choice of threshold is not made subjectively. With this approach, each method is given the chance to form a network at a comparable level of density, so that the comparison between methods becomes more objective and is not biased by a manually chosen threshold. The use of the LCC further strengthens the validity of the evaluation. By focusing the analysis on the largest connected component, the modularity computed genuinely represents the community structure of the main network. A mean node coverage of 97.17% shows that most households remain within the analysis, so that the LCC does not significantly reduce the representativeness of the data.

Based on the modularity results, the Jaccard Index can be recommended as the primary method for building village welfare networks, especially when the data have been represented in categorical or binary form. It is better able to capture substantive similarity because it computes the overlap of attributes genuinely shared by two households. Cosine Similarity can serve as an alternative when the goal is to read similarity in the direction of welfare profiles useful when the general orientation across attributes matters more than direct attribute similarity. Pearson Correlation can serve as a comparative method for testing whether the network structure remains stable when relationships are formed on the basis of aligned variation patterns. Taken together, these findings reinforce that the network construction stage must be treated as a primary

methodological concern: the choice of similarity measure, the thresholding mechanism, and the use of the LCC all determine network quality, the legibility of communities, and the validity of any subsequent interpretation.

At the same time, these recommendations are conditional. The robustness analyses above show that the ranking of methods can change under a different graph-construction scheme (Section 3.10) and that modularity measures structural quality rather than policy validity. The Jaccard Index should therefore be read as the most suitable measure for the specific combination of binary welfare representation and similarity thresholding used here, not as a universally optimal choice; external validation of the detected communities against welfare or poverty categories remains necessary before any policy use. It should also be acknowledged that part of the observed modularity may reflect the structure of the one-hot encoded binary feature space itself: discretizing continuous welfare scores into a limited set of categories produces groups of households that share identical category combinations, which similarity-based edges then reinforce. A modularity around 0.50 therefore indicates clear structural separability within this constructed representation, not proof that the resulting groups coincide with externally recognized welfare or poverty categories. This concern is consistent with a broader literature on the principled filtering and thresholding of weighted networks, in which the retained structure depends strongly on the construction rule adopted [34].

4. CONCLUSION

This study shows that the edge construction method exerts an important influence on the quality of the Village Welfare Index network structure. The three methods compared Cosine Similarity, the Jaccard Index, and Pearson Correlation are all able to produce analyzable networks, with a 100% success rate and an average LCC node coverage of 97.17%; the combination of automatic thresholding and Largest Connected Component extraction thus proves capable of keeping the network stable, neither too dense nor too sparse. Although the three methods yield relatively close modularity values, the Jaccard Index performs most consistently, with the highest mean modularity at 0.5074 and the best result in 10 of the 14 villages, which suggests rather than proves that edge construction based on attribute overlap is better suited to representing welfare proximity between households when the data are represented in categorical or binary

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

form; a Friedman test (chi-square = 8.71, p = 0.013) with Wilcoxon post-hoc comparisons confirmed that Jaccard's advantage over Cosine and Pearson is statistically significant, though the effect is specific to the tested representation and to the automatic-thresholding construction scheme. Cosine Similarity remains relevant for reading similarity in the direction of welfare profiles, and Pearson Correlation is useful as a comparative approach; methodologically, this study underscores that edge construction is a fundamental stage in welfare-data-based Social Network Analysis, one that determines the quality of the community structure. These findings are methodological in nature and do not yet evaluate policy impact. Their scope is bounded by several choices: the discretization and one-hot encoding of the welfare indicators, the grid-based automatic thresholding, the exclusive use of modularity as the evaluation criterion, and the stochastic nature of Louvain community detection. The study also does not validate the detected communities against external references such as official welfare or poverty categories, geographic hamlets (dusun), or policy outcomes; robustness analysis further shows that under a k-nearest-neighbor construction Pearson Correlation, not Jaccard, attains the highest modularity, underscoring that no single measure is universally optimal. Before any policy use, therefore, the detected welfare groups must be validated externally—against official welfare or poverty classifications and together with village experts. Building on this, future research could pursue three focused directions: validating the clusters against ground-truth welfare classifications and local socio-economic patterns; connecting the constructed networks to downstream analyses such as centrality, assortativity, and network-based policy auditing; and comparing the similarity-based networks with geographically defined networks (for example, hamlet/dusun proximity) and with networks built from real social ties such as kinship or interaction.

ACKNOWLEDGMENT

The authors thank the Master's Program in Computer Science, School of Data Science, Mathematics, and Informatics, IPB University, for its academic support in carrying out this research. The authors also extend their appreciation to all those who supported the provision of data and the development of the study's methodological framework.

REFERENCES

[1] United Nations Development Programme (UNDP), *Human Development Report 1990: Concept and Measurement of Human Development*. New York, NY, USA: Oxford Univ. Press, 1990.

[2] S. Alkire, "Dimensions of human development," *World Dev.*, vol. 30, no. 2, pp. 181–205, 2002, doi: 10.1016/S0305-750X(01)00109-7.

[3] Badan Pusat Statistik, *Indeks Pembangunan Desa 2018*. Jakarta, Indonesia: Badan Pusat Statistik, 2019.

[4] S. Sjaf, Sampean, A. A. Arsyad, L. Elson, A. R. Mahardika, L. Hakim, S. A. Amongjati, R. Gandi, Z. A. Barlan, I. M. G. Aditya, S. A. B. Maulana, and M. R. Rangkuti, "Data Desa Presisi: A new method of rural data collection," *MethodsX*, vol. 9, Art. no. 101868, 2022, doi: 10.1016/j.mex.2022.101868.

[5] N. Couldry and U. A. Mejias, *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford, CA, USA: Stanford Univ. Press, 2019, doi: 10.1515/9781503609754.

[6] N. Couldry and A. Powell, "Big Data from the bottom up," *Big Data Soc.*, vol. 1, no. 2, 2014, doi: 10.1177/2053951714539277.

[7] A. Majeed and I. Rauf, "Graph theory: A comprehensive survey about graph theory applications in computer science and social networks," *Inventions*, vol. 5, no. 1, Art. no. 10, 2020, doi: 10.3390/inventions5010010.

[8] F. Cacheda, D. Fernandez, F. J. Novoa, and V. Carneiro, "Early detection of depression: Social network analysis and random forest techniques," *J. Med. Internet Res.*, vol. 21, no. 6, Art. no. e12554, 2019, doi: 10.2196/12554.

[9] T. Adali and A. Ortega, "Applications of graph theory [Scanning the Issue]," *Proc. IEEE*, vol. 106, no. 5, pp. 784–786, May 2018, doi: 10.1109/JPROC.2018.2820300.

[10] S. B. Guerrero-Ocampo and J. M. Díaz-Puente, "Social network analysis uses and contributions to innovation initiatives in rural areas: A review," *Sustainability*, vol. 15, no. 18, Art. no. 14018, 2023, doi: 10.3390/su151814018.

[11] A. Levy, B. R. Shalom, and M. Chalamish, "A guide to similarity measures and their data science applications," *J. Big Data*, vol. 12, no. 1, Art. no. 188, 2025, doi: 10.1186/s40537-025-01227-1.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[12] S. Ghosh, N. Das, T. Gonçalves, P. Quaresma, and M. Kundu, "The journey of graph kernels through two decades," *Comput. Sci. Rev.*, vol. 27, pp. 88–111, 2018, doi: 10.1016/j.cosrev.2017.11.002.

[13] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech. Theory Exp.*, vol. 2008, no. 10, Art. no. P10008, 2008, doi: 10.1088/1742-5468/2008/10/P10008.

[14] Y. Lee, Y. Lee, J. C. Seong, A. Stanescu, and C. S. Hwang, "A comparison of network clustering algorithms in keyword network analysis: A case study with geography conference presentations," *Int. J. Geospat. Environ. Res.*, vol. 7, no. 3, pp. 1–16, 2020.

[15] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, no. 2, Art. no. 026113, 2004, doi: 10.1103/PhysRevE.69.026113.

[16] X. Cheng, C. Yang, Y. Zhao, Y. Wang, H. Karimi, and T. Derr, "A comprehensive analysis of social tie strength: Definitions, prediction methods, and future directions," *arXiv preprint arXiv:2410.19214*, 2024, doi: 10.48550/arXiv.2410.19214.

[17] F. Montes, R. C. Jimenez, and J.-P. Onnela, "Connected but segregated: Social networks in rural villages," *J. Complex Netw.*, vol. 6, no. 5, pp. 693–705, 2018, doi: 10.1093/comnet/cnx054.

[18] P. Li, J. Yu, J. Liu, D. Zhou, and B. Cao, "Generating weighted social networks using multigraph," *Physica A*, vol. 539, Art. no. 122894, 2020, doi: 10.1016/j.physa.2019.122894.

[19] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annu. Rev. Sociol.*, vol. 27, pp. 415–444, 2001, doi: 10.1146/annurev.soc.27.1.415.

[20] M. T. Rivera, S. B. Soderstrom, and B. Uzzi, "Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms," *Annu. Rev. Sociol.*, vol. 36, pp. 91–115, 2010, doi: 10.1146/annurev.soc.34.040507.134743.

[21] M. Suyudi, Sukono, and A. T. Bon, "Theoretical and simulation approaches centrality in social networks," in *Proc. 11th Annu. Int. Conf. Ind. Eng. Oper. Manag. (IEOM)*, Singapore, Mar. 7–11, 2021, pp. 4093–4101, doi: 10.46254/AN11.20210734.

[22] M. E. J. Newman, "Assortative mixing in networks," *Phys. Rev. Lett.*, vol. 89, no. 20, Art. no. 208701, 2002, doi: 10.1103/PhysRevLett.89.208701.

[23] U. von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007, doi: 10.1007/s11222-007-9033-z.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[24] M. Maier, M. Hein, and U. von Luxburg, "Optimal construction of k-nearest-neighbor graphs for identifying noisy clusters," *Theor. Comput. Sci.*, vol. 410, no. 19, pp. 1749–1764, 2009, doi: 10.1016/j.tcs.2009.01.009.

[25] M. Á. Serrano, M. Boguñá, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 106, no. 16, pp. 6483–6488, 2009, doi: 10.1073/pnas.0808904106.

[26] J. Olivier, W. D. Johnson, and G. D. Marshall, "The logarithmic transformation and the geometric mean in reporting experimental IgE results: What are they and when and why to use them?," *Ann. Allergy Asthma Immunol.*, vol. 100, no. 4, pp. 333–337, 2008, doi: 10.1016/S1081-1206(10)60595-9.

[27] A. Mahmoudi, M. R. Yaakub, and A. Abu Bakar, "A new method to discretize time to identify the milestones of online social networks," *Soc. Netw. Anal. Min.*, vol. 8, no. 1, Art. no. 34, pp. 1–20, 2018, doi: 10.1007/s13278-018-0511-4.

[28] M. Fernández, H. Kirchner, and B. Pinaud, "Labelled port graph—A formal structure for models and computations," *Electron. Notes Theor. Comput. Sci.*, vol. 338, pp. 3–21, 2018, doi: 10.1016/j.entcs.2018.10.002.

[29] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975, doi: 10.1145/361219.361220.

[30] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytol.*, vol. 11, no. 2, pp. 37–50, 1912, doi: 10.1111/j.1469-8137.1912.tb05611.x.

[31] K. Pearson, "Note on regression and inheritance in the case of two parents," *Proc. R. Soc. Lond.*, vol. 58, pp. 240–242, 1895, doi: 10.1098/rspl.1895.0041.

[32] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: Guaranteeing well-connected communities," *Sci. Rep.*, vol. 9, Art. no. 5233, 2019, doi: 10.1038/s41598-019-41695-z.

[33] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 104, no. 1, pp. 36–41, 2007, doi: 10.1073/pnas.0605965104.

[34] F. Radicchi, J. J. Ramasco, and S. Fortunato, "Information filtering in complex weighted networks," *Phys. Rev. E*, vol. 83, no. 4, Art. no. 046101, 2011, doi: 10.1103/PhysRevE.83.046101.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

APPENDIX A. Per-Village Results for Each Edge Construction Method

Table A1 consolidates, for each of the 14 villages, the network size after Largest Connected Component (LCC) extraction, node coverage, edge density, and edge count, together with the modularity obtained under each of the three edge construction methods. The LCC node count, density, and edge count are matched across methods at the automatically selected thresholds (Cosine = 0.3, Jaccard = 0.2, Pearson = 0.3 in every village; see Table 12). For each village, the highest modularity across the three methods is shown in bold.

Table A1. Per-village network summary and modularity by method

Village	LCC nodes	Coverage	Density	Edges	Mod. (Cosine)	Mod. (Jaccard)	Mod. (Pearson)
Bumi Ayu	179	92.7%	0.0358	570	**0.6307**	0.6303	0.6289
Curup	495	98.6%	0.0223	2,729	0.5278	0.5305	**0.5356**
Harapan Jaya	681	99.6%	0.0373	8,633	0.4549	**0.4585**	0.4500
Lunas Jaya	315	99.1%	0.0422	2,087	0.4532	**0.4592**	0.4542
Modong	418	99.3%	0.0276	2,408	0.5090	**0.5138**	0.5134
Muara Dua	210	91.3%	0.0288	631	0.5787	0.5783	**0.5788**
Muara Sungai	190	90.5%	0.0315	565	0.5435	**0.5491**	0.5423
Pandan	842	98.7%	0.0313	11,092	0.4818	**0.4839**	0.4826
Raja	775	100.0%	0.0481	14,430	0.4302	**0.4421**	0.4293
Sedupi	534	98.7%	0.0271	3,856	0.5078	**0.5096**	0.5082
Sukaraja (Tanah Abang)	314	99.7%	0.0355	1,746	0.5220	**0.5282**	0.5226
Tanah Abang Jaya	873	99.7%	0.0525	19,990	**0.4115**	0.4066	0.3999
Tanah Abang Utara	756	99.6%	0.0298	8,518	0.4640	**0.4752**	0.4673
Tanjung Dalam	213	93.0%	0.0413	932	0.5279	**0.5377**	0.5266

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

APPENDIX B. Number of One-Hot Encoded Binary Features per Village

Because the five Village Welfare Index dimensions are discretized and then one-hot encoded, the final number of binary features can differ across villages according to how many distinct category values occur in each village. Table B1 reports, for the main analysis (integer rounding), the number of one-hot categories contributing from each of the five dimensions (f_A–f_E) and the resulting total binary feature vector length per village.

Table B1. One-hot encoded binary features per village at integer rounding (main analysis)

Village	Households	f_A	f_B	f_C	f_D	f_E	Total features
Bumi Ayu	193	37	36	34	49	36	192
Curup	502	68	40	41	64	56	269
Harapan Jaya	684	60	35	39	59	50	243
Lunas Jaya	318	41	30	36	47	47	201
Modong	421	48	27	43	57	46	221
Muara Dua	230	54	28	38	46	43	209
Muara Sungai	210	36	28	39	46	42	191
Pandan	853	62	38	41	57	66	264
Raja	775	45	36	43	65	54	243
Sedupi	541	50	35	44	51	55	235
Sukaraja (Tanah Abang)	315	41	25	34	49	53	202
Tanah Abang Jaya	876	56	35	44	65	57	257
Tanah Abang Utara	759	59	45	43	62	58	267
Tanjung Dalam	229	45	24	33	45	45	192

The feature count grows as the welfare scores are rounded more finely, because finer rounding produces more distinct categories. Table B2 reports the total number of binary features per village under the three discretization levels examined in the robustness analysis (Section 3.9); the values in the integer column coincide with the totals in Table B1.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Table B2. Total number of one-hot binary features per village by rounding precision (cf. Section 3.9)

Village	Integer (0 decimal)	1 decimal	2 decimals
Bumi Ayu	192	297	301
Curup	269	519	539
Harapan Jaya	243	512	551
Lunas Jaya	201	344	361
Modong	221	393	400
Muara Dua	209	368	375
Muara Sungai	191	326	335
Pandan	264	601	635
Raja	243	534	559
Sedupi	235	451	479
Sukaraja (Tanah Abang)	202	340	353
Tanah Abang Jaya	257	596	649
Tanah Abang Utara	267	597	648
Tanjung Dalam	192	306	312