Analysis of Facebook Network

Tianfeng Liao
12212661

Chen Ma
12112941

Kexin Liu
12111651

***Abstract*—This report examines the Facebook network, consisting of 4039 nodes and 88234 edges, and confirms its scale-free structure with a power law degree distribution adjusted to γ= 2.03. The network exhibits assortative mixing with a coefficient of 0.0636. Community detection using the Louvain algorithm achieves a modularity of 0.835, improved to 0.839 with the Ego-Splitting Framework, reflecting Facebook's ego-centered nature. Information spread, modeled with SIR, reveals rapid dissemination facilitated by network bridges and structural holes. Comparing to a degree-preserve null model shows reduced clustering and different assortativity, highlighting the original network's unique social clustering and path complexity.**

I. Basic information

Facebook is a global social networking platform founded in 2004. It enables users to connect with friends, family, and colleagues, share updates, photos, and videos, and engage in a variety of online communities and groups.

Facebook data is collected from survey participants using this Facebook app. It contains **4039 nodes** and **8823 links**. The basic information of this network is showed in table I and the network is showed in fig 1

number of nodes	4039
number of edges	88234
average degree	43.69
diameter	8
average shortest path length	2.96
average clustering coefficient	0.6055
largest component	4039

Top 5 hubs	
node	degree
107	1045
1684	792
1921	755
3437	547
0	347

TABLE I: Basic information

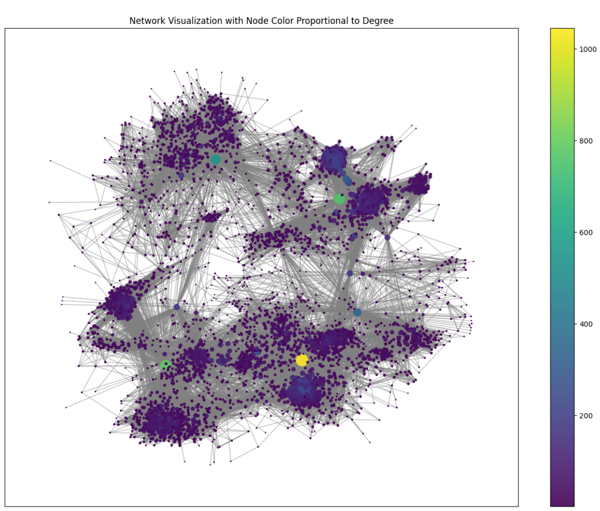

Fig. 1: Facebook network

II. Degree distribution

From the degree distribution in fig 2 of the network, we see it is a scale-free network.

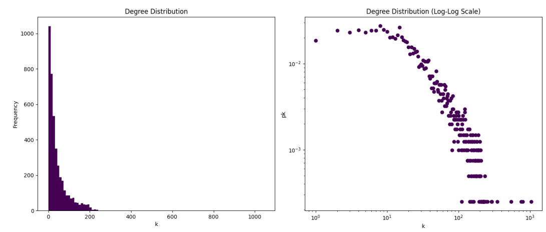

Fig. 2: Degree distribution

We fit the network with power law distribution, which is:

$$p_k = k^{-\gamma}$$

As showed in fig 3, $\gamma = 5.99$

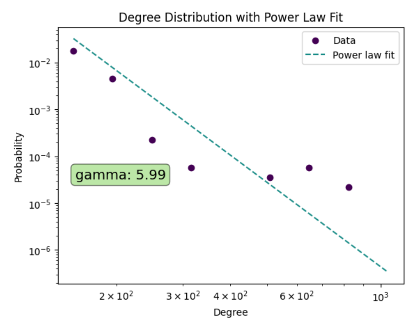

Fig. 3: Power law

However, because of the **Low-degree saturation** and **High-degree cutoff** properties of real world network.

- Low-degree saturation is a common deviation from the power-law behavior. Its signature is a flattened p_k for $k < k_{sat}$. This indicates that we have fewer small degree nodes than expected for a pure power law. This is because that unlike the theoretical network, small degree nodes have initial attractiveness, the degree distribution changes into $p_k = C(A + k)^{-\gamma}$. Therefore, initial attractiveness induces a small-degree saturation for $k < A$.
- High-degree cutoff appears as a rapid drop in p_k for $k > k_{cut}$, indicating that we have fewer high-degree nodes than expected in a pure power law. This limits the size of the largest hub, making it smaller than predicted by $k_{max} = k_{min} N^{\frac{1}{\gamma - 1}}$. In this case, it is reasonable because

it is impossible for one people to have relationship with almost all community members.

Thus, we change the fit of degree distribution into:

$$\tilde{p_k} = p_k \times exp(\frac{k + k_{sat}}{k_{cut}})$$

Where we choose $k_{sat} = 10$, $k_{cut} = 500$.

Then Low-degree saturation and high degree cutoff trends are vanished in degree distribution in fig 4 and power law distribution coefficient γ changes to $\gamma = 2.03$ as showed in fig 5 and is more reasonable and suitable for real world.

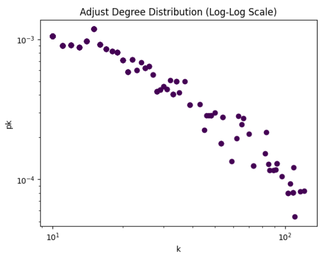

Fig. 4: Adjusted degree distribution

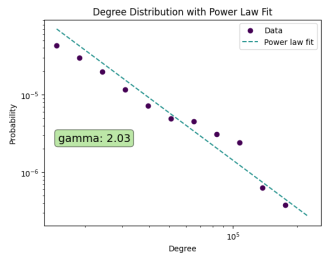

Fig. 5: Adjusted power law

III. Degree correlation

From heat map of degree correlation in fig 6, we see that the network is an assortative network and has **degree assortativity** equal to 0.0636.

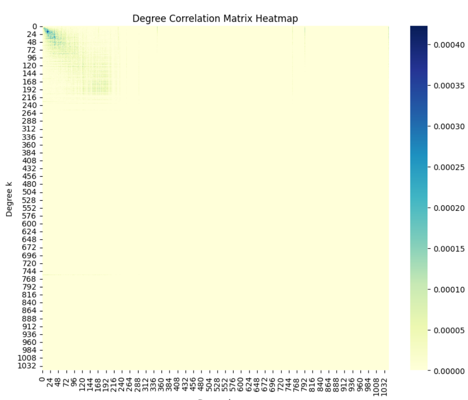

Fig. 6: Heat map of degree correlation

A. Degree correlation function

Then we derive degree correlation function relation and we want fit it with $k_{nn} = a\mu^k$

However, we see that there is a decreasing trend before k=10 and sudden cutoff after k=500, which leads to underfitting of degree correlation function. These may because some low degree nodes only have relationship with some high degree nodes like we see in **Friends Paradox** and cutoff at high degree has the same reason like we put forth. Thus, we only use nodes with degree between 10 and 500, and we have $a = 10.575, mu = 0.497$.

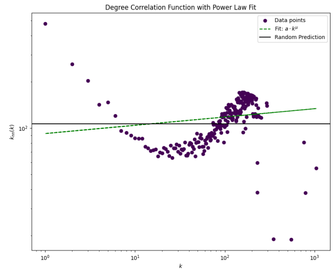

Fig. 7: Degree correlation function

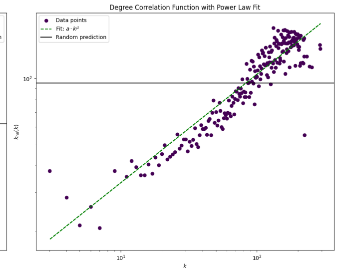

Fig. 8: Degree correlation function with cutting

B. Degree preserving randomization

We use both **Degree Preserving Randomization with Simple Links (R-S)** and **Degree Preserving Randomization with Multiple Links (R-M)**, from fig 9, we see that assortative nature disappears which indicates that the assortative correlations of the collaboration network is not linked to its scale-free nature.

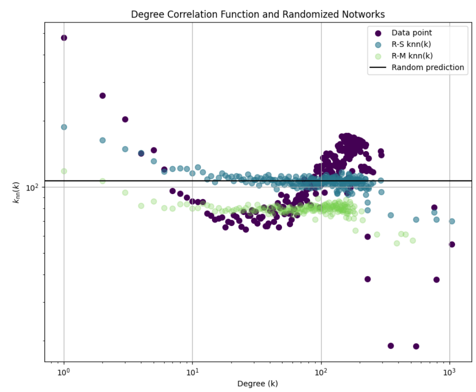

Fig. 9: Randomization

IV. Community Detection

Facebook contains various communities and interest groups. Therefore, it is natural for us to explore the communities within the Facebook network.

We will employ different algorithms to investigate these communities, with the aim of identifying the algorithm that maximizes modularity. It is understood that algorithms maximizing modularity may sometimes forcefully merge two communities connected by only one edge; however, we will not consider this scenario for the time being.

A. Infomap Algorithm

First, the Infomap algorithm will be used. This algorithm is based on information theory and detects network structures by minimizing the average length of information coding.

The **modularity** of this algorithm is **0.706**. Figure 10 is the community structure of Facebook network detected by Infomap Algorithm. Figure 11 is the community size distribution. In the communities detected by this algorithm,

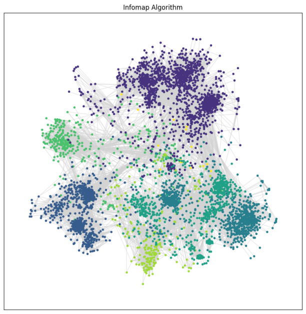

Fig. 10: Community Structure detected by Infomap

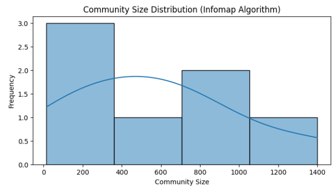

Fig. 11: Community Size Distribution detected by Infomap

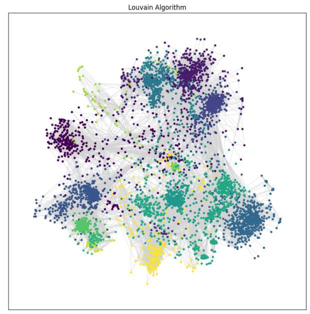

Fig. 14: Community Structure detected by Louvain

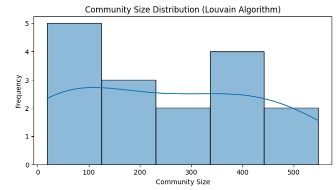

Fig. 15: Community Size Distribution detected by Louvain

small communities are few and small, while large communities contain more than one-third of the total number of nodes in the network. We believe that this model may have merged some small communities into larger ones. Overall, this is not a very ideal algorithm.

B. Label Propagation Algorithm

The Label Propagation Algorithm infers the community structure of the entire network based on the local information of nodes within the network.

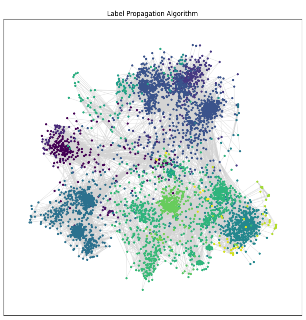

Fig. 12: Community Structure detected by LPA

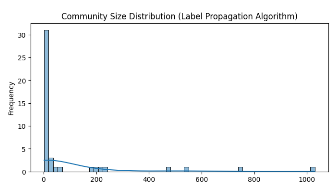

Fig. 13: Community Size Distribution detected by LPA

The **modularity** of this community detected by LPA is **0.736**. From the community size distribution plot, we can see that there are many small community and there exist one huge community. The number of medium-sized communities is very small. We speculate that this algorithm has merged most medium-sized communities into one large community. This may not accurately reflect the real communities in Facebook, so we will not choose the communities detected by this algorithm for further analysis.

C. Louvain Algorithm

The core of this algorithm is the optimization of modularity. The Louvain algorithm is a community detection method used to extract non-overlapping communities from large networks. It was introduced by researchers from the University of Louvain. From the distribution of community sizes, we can see that the algorithm divides the entire graph into several communities of varying sizes. There are fewer large communities and more small communities, which aligns with our expectations for the Facebook network. The modularity of the communities identified by this algorithm is 0.835, showing a significant improvement over previous algorithms.

D. Ego-Splitting Framework

Next, we attempt to find an algorithm suitable for our network, based on the characteristics of the Facebook network. Our Facebook network is an ego-centered network. The Ego Network is a concept in social network analysis that refers to a subset of the network centered around a specific individual (called the "ego"), including the ego and their directly connected social network members (called "alters"). In this network, the "ego" is the central node, while the "alters" are other nodes directly connected to the ego.

Based on the characteristics of the ego-centered network, we use an algorithm within the Ego-Splitting Framework. This framework is an improvement of the Louvain algorithm for ego-centered networks.

Next, we will use this algorithm to perform community detection.

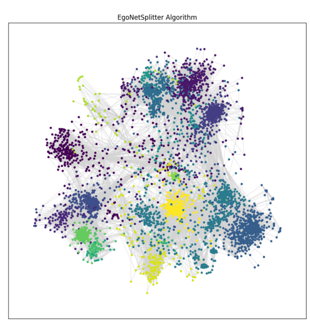

Fig. 16: Community Structure detected by Ego-Splitting Framework

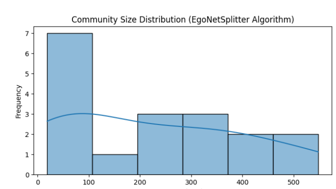

Fig. 17: Community Size Distribution detected by Ego-Splitting Framework

The modularity obtained by this algorithm is 0.839, which is an improvement over the Louvain algorithm. The community size distribution identified by this algorithm also aligns better with the characteristics of the Facebook network. Due to the lack of certain attributes (such as node and edge attributes), we are unable to further explain the existence of these communities.

Based on the detected communities, we will discuss the spread of rumors within communities in the next section.

V. The Spreading Phenomenon

Since we have discussed the community behavior in the former section, here we are going to explore how information can be spread through the Facebook social network. First let us go through some basic network structures related to the spreading phenomenon, including bridges, structural holes and potential links.

A. Bridges

An edge joining two nodes A and B in the graph is considered a bridge, if deleting the edge would cause A and B to lie in two different components. In detail, an edge joining two nodes and in a graph is a local bridge, if its endpoints and have no friends in common. Very importantly, an edge that is a bridge is also a local bridge.

First we need to check if there are any bridges in this network and the result is true. And there are 75 bridges and 78 local bridges, which can be used to spread information between two clusters, facilitating the rapid dissemination of new ideas and behaviors across the network and serving as vital conduits in scenarios such as viral marketing or public health campaigns.

Here we mark the bridges with red lines denoting to normal bridges and green to local bridges in Figure 18. The result matches our intuitions as information is able to easily spread between social network communities since there are a large number of bridges.

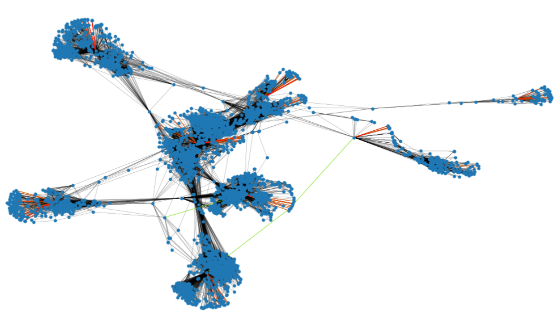

Fig. 18: Bridges of the Network

B. Structural Holes

Local bridges are particularly critical for the dissemination of new information as they often represent the only direct link between highly disparate social groups. Also, the concept of structural holes are similar to local bridges' way: gaps between groups in a social network where there is no direct communication.

Nodes that bridge these gaps can access diverse information and resources from different parts of the network. Typically, nodes in structural holes are those that connect distinct dense groups or communities, gaining access to a wealth of information and resources by bridging different parts of the network. Here we mark this kind of nodes red in the network in Figure19.

Surprisingly, it can be noticed that these edges appear at the edges of each cluster. This is because these nodes may be connecting smaller groups or individual outliers far from the center, causing them to appear at the edges visually. This positioning allows them to bridge communities at a distance, which might not connect directly without these nodes. Hence, structural holes are not always showing prominently between large clusters. Sometimes, they are located in seemingly inconspicuous positions, linking groups that are less active or prominent on the surface. Understanding the placement

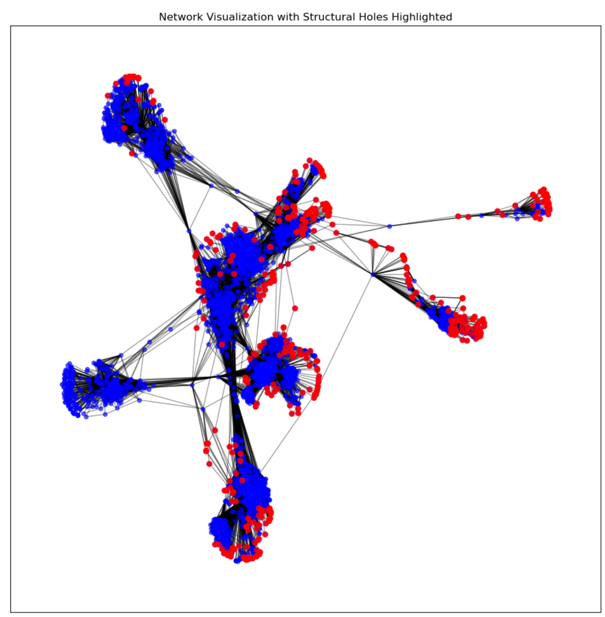

Fig. 19: Structural Holes of the Network

of structural holes is crucial for uncovering the dynamics of information flow and influence within the network.

C. Potential Links

The formation of potential links can alter the pathways through which information travels, enhancing the network's connectivity and potentially its efficiency in spreading processes. We use common neighbors to denote the chance of forming a link in the future: the more neighbors two nodes have in common, the higher the probability of future links. With the number of common nodes marked in red, we select top 5 pairs in the network and the visualization is as follows:

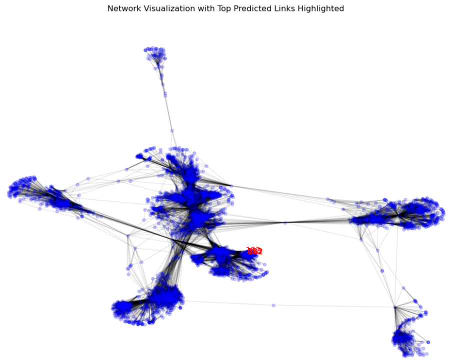

Fig. 20: Potential Links

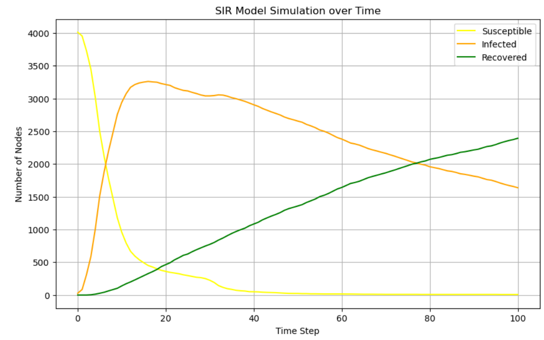

Fig. 22: SIR Simulation Over Time

It can be concluded that such pairs are within a extremely dense cluster and they are pretty close.

D. SIR Modeling

Now we are ready for the SIR model, in which S, I, and R represent three distinct states: Susceptible, Infected, and Recovered. This model is commonly used to simulate the spread of infectious diseases but can also be applied to the dissemination of information in social networks. In the information spreading, S denotes to individuals who have not yet received the information but are susceptible to becoming 'infected', I to the individuals who have received the information and can potentially spread it to other susceptibles while R to who no longer spread the disease or information, possibly because they have recovered or lost interest in the information. Initially, we set initial infected nodes as 10, infection probability as 0.05, recovery rate as 0.01 and we iterate 100 steps. Here we use two plots to illustrate the result:

The simulation starts with a vast majority of the population

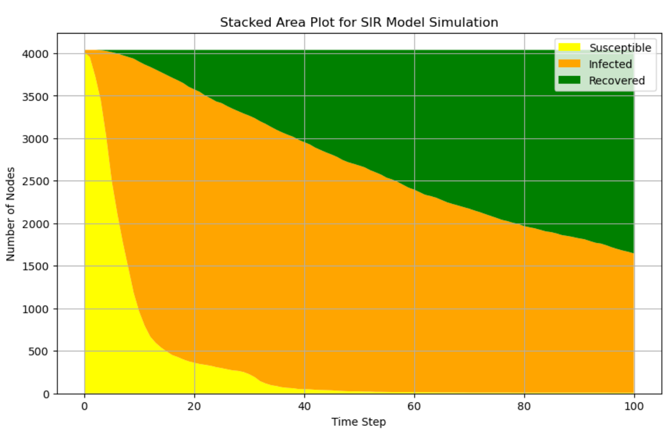

Fig. 21: Stacked Area Plot for SIR model

in the susceptible state (yellow), a small proportion infected (orange), and none or very few recovered (green).

The infected population (orange) grows quickly initially, indicating a rapid spread of the infection. This could suggest high initial contact rates or a highly contagious condition within a densely connected network. The curve for the infected population then peaks and starts declining, which suggests that the infection rate slows down, due to a decreasing number of susceptible individuals or effective recovery rates.

The recovery group grows with a very steady rate and gradually the recovery curve overtakes the infection curve, showing that the recovery rate outpaces new infections over time.

Toward the end of the simulation, the majority of the population appears to have recovered (green), with very few remaining susceptible or infected. This suggests that the spreading process may be nearing its end within this network.

The result corresponds to our intuitions since information spreads extremely fast and widely through the internet social work(very few people remain susceptible at the iteration end). And this is why the internet era is labeled as "information explosion". Though the major of the news will eventually be abandoned form people's memories(not again interested in the specific information), there are still traces on the internet, even the social network.

VI. The Degree Preserve Null Model

We have discussed the spreading phenomenon and know the social network's properties in the former sections. Now we are going to analyze more on these properties.

The Null Models are used as a benchmark to understand the significance of the observed network structures. Degree Preserve Null Models typically focus on reconfiguring a network while maintaining each node's degree. This approach allows us to differentiate the network properties emerging from the network's degree distribution alone from those properties resulting from the network's specific structure.

After generating a random network while maintaining the degree distribution, we first remove self-loops in the network before directly showing the null model's visualization.

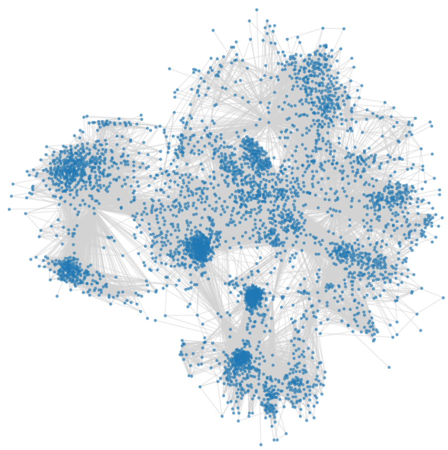

Fig. 23: The Original Network

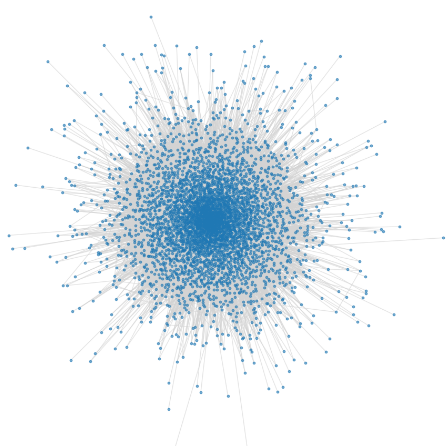

Fig. 24: The Null Model

The original network graph displays a complex community structure, including several distinct communities. These communities are connected by fewer bridges, showing the typical modular features of a social network, reflecting the formation of real-world social circles where people tend to form close connections within small groups sharing common interests.

The null model graph shows a network with a uniform connection pattern, where no clear clustering or community structures are evident. All nodes appear almost equally likely to connect randomly, forming a radial layout. This structure results from the null model disrupting the original network's edges while preserving the nodes' degree distribution, leading to random reconnections of edges.

Now we calculate some basic metrics of the null model to check our graph and the result is in the following table:

TABLE II: Network Characteristics Comparison

Metric	Original Network	Null Model
Clustering Coefficient	0.6055	0.0525
Average Path Length	3.6925	2.6264
Degree Assortativity	0.0636	-0.0245

We confirm that the null model is still connected, suggesting robust connectivity under the degree distribution despite randomization.

The significant drop in the clustering coefficient from the original network to the null model suggests that the original network has a much higher level of local clustering, indicative of more closures and potentially small-world characteristics among the social network.

The shorter path length in the null model compared to the original network indicates that, while the degree distribution contributes to maintaining relatively short paths (small-world property), the specific connections in the original network might add complexity resulting in longer paths.

The original network shows a slight preference for nodes to connect with others that have similar degrees (positive assortativity), whereas the null model has a slight negative assortativity. This suggests that the network's original assortative mixing pattern is not solely due to its degree distribution but may be an inherent property of its specific structure of real-world network.