# Fairness Aware Influence

## Study of relationship between node network properties and FAI in complex networks

Varnika Srivastava


Delft University of Technology

TU Delft

# Fairness Aware Influence

## Study of relationship between node network properties and FAI in complex networks

by

Varnika Srivastava

| Student Name | Student Number |
|---|---|
| Srivastava | 5534178 |


| | |
|---|---|
| Supervisor: | Dr. Huijuan Wang |
| Project Duration: | November, 2022 - November, 2023 |
| Faculty: | Faculty of Computer Science, Delft |



TU Delft

# Preface

Since my bachelors, I had an interest for network science. I expressed this research interest to the professor of my complex networks (MDACN) course here in TU Delft, who is now my supervisor for this thesis.

She proposed a problem statement to study the fraction of infections in each community. Upon conducting a review of exisiting literature, I found there were similar problems related to influence maximization, community diversity, etc, but none of them had the same problem statement.

Thus, this thesis titled - ”Fairness Aware Influence”, introduces single spreader fairness aware influence(FAI) as a novel concept, where in addition to a given seed node’s prevalence, we study the variation of fraction of infections in all available communities.

The main objective is to study the relationship of FAI with network properties of the node such as degree, community size, etc. Existing centrality metrics are not designed for FAI, so two classes of new metrics are proposed which incorporate different network and community properties. These new metrics outperform the existing ones at $\lambda_c$ given they consider a sufficiently large neighborhood of the seed node in question. The dynamics of FAI changes at higher infection levels, and we find that the performance of the proposed metrics are not stable at these infection levels.

Dr.Johan Dubbeldam, along with my supervisor Dr.Wang, has consented to be a member of the thesis committee. Dr.Wang has my gratitude for her guidance without which this thesis would not have come to fruition. I would like to acknowledge the encouragement and help I received from my supervisor’s PhD students - Shilun, Li, Omar and Alberto, and my peers Mathieu and Yuhui. Last, but not the least, I would like to thank my friends and family for their constant support throughout my thesis.

*Varnika Srivastava*
*Delft, November 2023*

# Summary

This thesis introduces the concept of Fairness Aware Influence (FAI), which is dependent on prevalence and fairness. Prevalence of a given seed node is the number of infected nodes. Fairness can be defined differently based on the application or problem statement. In this case, the fairness is defined as the variation in the fraction of infections in all the communities. This variation is measured using standard deviation (SD). A lower SD corresponds to better fairness. FAI for a given seed node is defined as the ratio of prevalence to fairness, where a higher FAI score corresponds to higher prevalence and lower SD.

The primary objective of this thesis is to study how network properties such as degree and community size, relate with FAI. Network properties are measured using centrality metrics, which are categorized into two types. The first type, referred to as "simple" or classic centrality metrics, do not account for community information. The second type, known as community-aware centrality metrics, incorporate community information but were not originally designed for FAI. These serve as baselines for ranking nodes in terms of FAI. Thus, two new classes of metrics are designed specifically for FAI in the attempt to perform better than the baselines.

Six real world networks are employed to evaluate the metrics. Local centrality and Community-Hub-Bridge are found to be good baselines in their respective categories, and the newly proposed metrics surpass the existing ones at the epidemic threshold $\lambda_c$. Additionally, a discussion is presented to compare and analyze these metrics, considering their performance under varying infection rates using an SIR infection spreading model.

# Contents

# 1

# Introduction

A complex network usually comprises various topological communities wherein a group of nodes tends to form more relations (links) with each other than nodes from other groups [18]. In complex networks that have community structure, information can spread to different extents in the communities given a set of seed nodes. This is undesirable in a scenario where information should be spread as fairly as possible to the different communities. For example, fairness is important in a social context where certain smaller (minority) communities tend to receive information much differently compared to other communities, as shown by Tsang et al. [25].

This thesis aims to study the *fairness and prevalence* of information spread in complex networks with strong community structure. Typically, information spread is modeled using various epidemic models such as SIR, SIS, etc, and the influence of a seed node is measured in terms of its prevalence i.e.; the number of nodes in the network infected in the stable state of the information spread. However, prevalence fails to consider the proportion of nodes infected in the various communities. Thus, a new method is needed to measure how the infection varies in the different communities. This leads to measuring the ”fairness” of a seed node. Fairness can thus be generally defined as the variation in the fraction of nodes infected among all the various communities present in the network.

The purpose of this research is to understand if the different communities can be fairly (equally) infected in addition to maximizing prevalence given a single seed node. In other words, we see if nodes can achieve high influence while being fairness-aware. This problem is termed as single-spreader Fairness Aware Influence(FAI).

Our central research question is - ”Does the FAI depend on network properties of the node?”. We study network properties to discern the factors influencing the FAI of a node. Such network properties can be used to rank nodes, as an estimation of the ranking of nodes based on FAI. While previous research has ranked nodes to identify nodes with the highest influence using established network centrality metrics such as degree centrality and closeness centrality, the concept of fairness in community infection is relatively new, and there are no known metrics in the literature designed to estimate FAI.

Classic centrality measures, while useful, do not account for community information, which is essential in measuring fairness. Although various community-aware metrics exist in literature that incorporate some community features [6, 7, 8, 9, 17, 26, 27], these are primarily designed for estimating prevalence, not fairness. Therefore, the main contribution of this research is the development of new community-aware metrics specifically tailored to estimate ranking of nodes in terms of FAI.

To this effect, two new classes of metrics are proposed. One is based on the average of community size of the neighboring nodes and itself. The other is based on estimating the FAI rank of a node by combining individual metrics designed to indicate either prevalence or fairness. The differentiation of nodes in relation to prevalence is achieved through the use of Local Centrality, while the assessment of fairness of a node is achieved using standard deviation of the fraction of nodes of each community present in its neighborhood. The metrics are evaluated using recognition rate to compare performance at ranking the top nodes, and kendall’s rank correlation to compare the overall performance across all nodes in the network. We find that the metric based on community size exceeds all others in terms of recognition rate, and has one of the highest rank correlation values.

The performance of these metrics are expected to change for different levels of infection. As the rate of transmission of infection increases, the prevalence and therefore, FAI are expected to change, potentially affecting how well these metrics estimate FAI rank of the nodes. Thus, we propose a second research question - ”How does the performance of the metrics change with increasing rate of transmission of infection?”.

The thesis is structured as follows. The existing literature is reviewed in chapter 2. The definitions of fairness and fairness aware influence are described in chapter 3. The methods and data used along with the implementation of the experiments are detailed in chapter 4. The results are presented and discussed in chapter 5. The conclusion and future work are discussed in chapter 6.

# 2

# Related Work

Fairness is a central theme in this research. Based on the literature study conducted, we can define fairness differently depending on the problem statement. For example, Ali et al. [1] define fairness of influence spread based on different ”groups” that are infected. The groups in a given network can be characterized by different social constructs such as gender, race, etc implying that nodes are of a certain ”type”. It also considers the time-critical aspect of information spread in a network. Thus, this thesis differs in two respects: there is no time-critical aspect of information diffusion involved and the groups are defined by communities and not various types of nodes.

Literature also delves into a term similar to fairness called diversity. Like fairness, diversity is also defined differently - either based on community [16] or based on the type of node [24].

Li et al. [16] propose the problem of community-diversified influence maximization (CDIM), which considers, in addition to prevalence, the number of communities to which the influence spreads. The CDIM problem deals with a given number of seed nodes and tries to maximize the CDIM objective with various algorithms. The FAI problem distinguishes itself from the CDIM problem by considering infection percentages in all communities, regardless of its activation by a given seed node. The CDIM problem only counts the number of activated communities.

Tang et al. [24] take the approach of diversifying seeds to reach an audience with more diverse backgrounds. The diversity stems from nodes belonging to various categories. Since it focuses on diversity and Influence maximization, this paper proposes a generalized objective function as a linear combination of prevalence and diversity. Using this general formula, a set of diversity measures can be constructed in combination with existing heuristics degree centrality and page rank to give diversified degree centrality and diversified page rank. Maximizing this objective then helps identify the most diverse set of seeds.

Considering the literature presented earlier, it can be concluded that this thesis presents a unique definition of fairness. It is not an optimization problem and instead focuses on network properties of a given seed node, including community information, to estimate Fairness Aware Influence.

Network properties of a node can be measured using centrality metrics. Bucur [3] uses classic centrality metrics of a node to predict its prevalence when this node is the single seed node of the spreading process. Centrality metrics (CMs) such as degree, betweenness, etc are plotted against prevalence, revealing a positive correlation for some graphs. However Bucur emphasizes that a single metric is often insufficient to predict prevalence, and a combination of them performs better, even across diverse graphs. In this study, we adopt a similar approach to examine the correlation between FAI and CMs. Our newly proposed metrics are designed to capture different node properties such as community size, number of neighbors, etc. However, we take a different perspective; we use each centrality metric to estimate the node’s rank in terms of its FAI instead of *predicting* FAI.

In addition to the CMs, there are community-aware centrality metrics (CACM) that consider community information. These are designed to capture community information that the classic CMs overlook. Rajeh et al. [21] compare seven CACMs in terms of their effectiveness in identifying prevalent nodes in various types of graphs, each with different degrees of community strength. In this study, we build upon this research by exploring these CACMs and evaluating their performance in ranking nodes in terms of FAI, comparing their efficacy to traditional CMs and the newly proposed metrics.

# 3

# Fairness Aware Influence

The problem statement of this research is to rank the nodes in a given network in terms of prevalence and fairness. It is essential to differentiate nodes with high influence (prevalence) and high fairness. Fairness Aware Influence (FAI) thus is a combination of two components - prevalence and fairness. These are defined as follows.

The prevalence $p$ of a seed node is the number of infected nodes in the steady state of the infection spreading process. However, prevalence does not consider the distribution of fraction of infections among the communities in the network. To measure this, a second component called fairness is defined.

Typically, fairness of infection (or information) spread can have multiple definitions depending on the context. In this case, fairness of infection spread is defined as the variation in fraction of infections among all the communities present in the given network. There are multiple ways to capture variation, such as range, standard deviation (SD) etc. Since the aim is to study infection in *all* communities, the range is not appropriate as it does not give us an idea of the distribution of infection other than the communities with maximum and minimum fraction of infection. Thus, intuitively, the difference of infections in communities can be captured appropriately through SD. The SD computes the deviation of the infection in a given community from the mean. A lower SD suggests higher fairness.

The mathematical formulation of fairness $f$ is as follows. Let $C$ be the set of all communities in the given network. $|C|$ is the total number of communities in the network. Let the fraction of infected nodes per community $c$ be $p_c$, the fraction of infected nodes in the network or prevalence be $p$. The average fractions of infections per community is $\mu_c$.

$$f = \frac{\sqrt{\Sigma_{c \epsilon C}(p_c - \mu_c)^2}}{|C|} \tag{3.1}$$

In addition to high prevalence, it is desirable to select a seed node with high fairness or a low $f$. Thus, one of the ways to define the term FAI ($F$) is :

$$F = \frac{p}{f} \tag{3.2}$$

FAI is large when the $p$ is large and $f$ is small (or fairness is high).

# 4

# Method

This section describes the methods used to conduct the experiments. The spreading process i.e.; the Susceptible-Infected-Recovered (SIR) model is introduced in 4.1, along with the calculation of epidemic threshold. The networks(data) used for the experiments are described in 4.2. The FAI is calculated for every node in the networks using the description in equation 3.2 based on the steady state of the SIR process on the given network.

Various centrality metrics that capture the network properties of a node to estimate the ranking of FAI are described in sections 4.3 and 4.4. These existing metrics are broadly categorized into two classes - ones that do not contain any community information and the ones that do.

Since these existing metrics are not designed specifically for FAI, two classes of new metrics are proposed. The design of the new metrics is described and motivated in 4.5. The methods used to evaluate the effectiveness of these metrics to rank nodes in terms of FAI are described in section 4.6. Finally, the experiment setup is described in 4.7.

## 4.1. SIR model and Epidemic Threshold

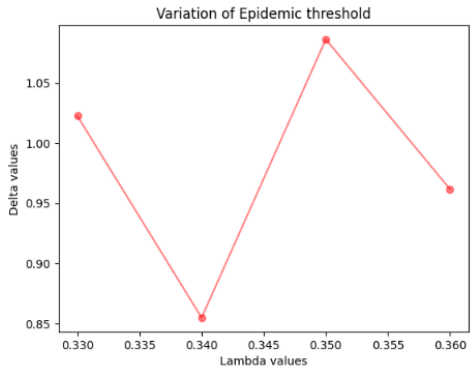
The SIR model is an infection diffusion model where nodes in the network can have three possible states - susceptible(S), infected(I), and recovered or removed(R). To start the infection-spreading process, a single node is chosen to be initially infected. This node is the seed node. Any node that comes into contact with an infected node becomes susceptible(S) to infection. Each infected node infects each of its susceptible neighbors independently with an infection rate $\beta$. Each infected node may recover(R) from the infection at a recovery rate $\gamma$. Both infection and recovery processes are independent Poisson processes. The infection-spreading process continues till there are no more infected nodes in the network. This state is called the steady state. The ratio of infection rate to recovery rate is called the effective transmission rate($\lambda$).

$$\lambda = \frac{\beta}{\gamma}$$

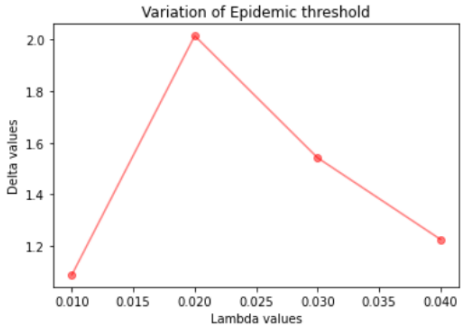
There exists a certain value of the effective transmission rate beyond which the infection reaches epidemic proportions. This value of $\lambda$ is called the epidemic threshold ($\lambda_c$). For $\lambda < \lambda_c$, the infection dies out rapidly [4]. Bucur [3] calculates epidemic threshold as the value of $\lambda$ at which the variability measure $\Delta$ is maximized. $\rho$ is the random variable of the outbreak size i.e.; the number of recovered nodes in the steady state of the infection spreading process for a given seed node, and operator $< . >$ represents the mean. The variability measure is computed as follows:

$$\Delta = \frac{\sqrt{< \rho^2 > - < \rho >^2}}{< \rho >} \tag{4.1}$$

We compute the FAI of the nodes at different transmission rates - $\lambda_c$, $2\lambda_c$ and $3\lambda_c$. The motivation for studying FAI at $\lambda_c$ is based on the coefficient of variation of FAI and is discussed further in 5.1.




**(a)** US Powergrid




**(b)** Facebook Politician Pages, $\lambda_c = 0.02$

**Figure 4.1:** Delta - Effective transmission rate graphs to determine $\lambda_c$. The $\lambda$ corresponding to the highest $\Delta$ value is the effective transmission rate $\lambda_c$.

## 4.2. Data

We consider Lancichinetti-Fortunato-Radicchi(LFR) benchmark graphs[14] and six real-world networks. The description of the network and other parameters and properties are elaborated in 4.2.1 and 4.2.2 respectively. For disconnected graphs the largest connected component was selected.

### 4.2.1. LFR graph

The LFR model generates networks with the degree of nodes and the community sizes following the power law distribution and with inherent community structure. One method to measure this community structure is to use modularity $Q$ which is a measure of how well the graph is divided into communities. LFR graphs tend to have high modularity. The communities are made readily available by the LFR graph generation algorithm.

We select the following controllable parameters - number of nodes, average degree, minimum and maximum number of nodes in a community (minC and maxC respectively), mixing parameter, and the power law exponent of the distributions of the node degree and size of the communities ($\tau_1$ and $\tau_2$ respectively). The mixing parameter $\mu$ of a node is the ratio of links to nodes outside its community to the total number of links of the node.

We consider number of nodes to be 1000 and average degree to be 4. The original paper [14] states that values of $2 <= \tau_1 <= 3$ and $1 <= \tau_2 <= 2$ are typically seen in most real-world networks. We therefore take $\tau_1 = 3$ and $\tau_2 = 2$. A set of values $\{0.1, 0.2, 0.3\}$ are considered for $\mu$. $\mu$ can take values between 0 and 1, but for $\mu$ > 0.3, the modularity of the graph becomes low. The minimum and maximum size of the community typically need to be specified in order for the graph generation algorithm to converge. It is difficult to choose suitable values for some of these parameters since there is no standard specified in related literature, and also, the values differ for different real-world networks. We consider the set of values $\{50, 100, 150, 200\}$ for $minC$ and $\{400, 450, 500\}$ for $maxC$. This graph can have self loops which should be removed before computing the graph properties such as modularity, mixing parameter, etc.

### 4.2.2. Real-world networks

To further test the metrics, six real world networks were chosen. These are Facebook Ego [15] combined network, Yeast Collins[20], Yeast Proteins [22], Facebook Politician Pages[22], LastFM Asia [15] and US Powergrid[13].

The communities were identified based on inherent network structure, using Louvain method of community detection [2]. The resolution parameter was set to $1e-100$ to favor larger sizes of communities. Without this specification, the number of communities tends to increase significantly for some graphs and this is not desirable. The six graphs were selected because each shows strong to moderate community structure as indicated by the high modularity values $Q$ in Table 4.1. The mixing parameter $\mu$ was also computed.

The basic graph properties along with the respective epidemic thresholds are summarized in Table

| Network | \|N\| | \|E\| | <k> | \|c\| | $\mu$ | Q | $\lambda_c$ |
|---|---|---|---|---|---|---|---|
| FacebookEgo | 4039 | 88234 | 43.69 | 12 | 0.02 | 0.73 | 0.01 |
| Yeast-Collins* | 1004 | 8319 | 16.57 | 71 | 0.11 | 0.72 | 0.02 |
| Yeast-proteins* | 1458 | 1948 | 2.73 | 256 | 0.19 | 0.68 | 0.17 |
| Facebook Politician Pages | 5908 | 41706 | 14.12 | 210 | 0.13 | 0.82 | 0.02 |
| Last-FM Asia | 7624 | 27806 | 7.29 | 526 | 0.18 | 0.73 | 0.04 |
| US Powergrid | 4941 | 6594 | 2.66 | 1222 | 0.29 | 0.63 | 0.35 |

**Table 4.1:** Properties of six real-world networks
$|N|$ is the number of nodes, $|E|$ is the number of edges, $< k >$ is the average degree, $|c|$ is the number of communities, $\mu$ is the mixing parameter, $Q$ is the modularity, and $\lambda_c$ is the epidemic threshold.
$*$ means that the largest connected component is used since the graph is disconnected

4.1. The epidemic threshold was calculated as described in chapter 4.1. The degree distribution of the networks are presented in A.1, and the community size distribution in A.2.

## 4.3. Centrality metrics

We consider the following centrality metrics in network science - degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, and local centrality.

Degree centrality of a node is the fraction of nodes in the graph it is connected to. Betweenness centrality of a node is defined as the fraction of the number of shortest paths between all node pairs that pass through it. Closeness Centrality of a node is the reciprocal of the average of the length of the shortest path to all other nodes in the network. It describes its efficiency of spreading information to other nodes [19]. Eigenvector centrality of a node is the corresponding element in the eigenvector associated with the largest eigenvalue of the adjacency matrix of the network. It captures the importance of a node based on its degree and the importance of other nodes it is connected to [23].

The computational complexity of betweenness and closeness centralities is high [19]. To overcome these performance issues, Chen [5] proposes a semi-local metric called local centrality ($C_L$). Chen shows that $C_L$ predicts prevalence in the SIR infection spreading process better than betweenness and degree centralities while being less computationally expensive than closeness and betweenness centralities. The formal definition of $C_L$ is as follows.

$$C_L(i) = \sum_{v \epsilon \Gamma(i)} Q(v)$$

where $Q(v) = \sum_{w \epsilon \Gamma(v)} N(w)$. The immediate neighbors of node $i$ are $\Gamma(i)$ and $N(w)$ is the number of nearest and next-nearest neighbours of node $w$.

## 4.4. Community-aware centrality metrics

While the community structure is inherent to many real-world networks [18], the centrality metrics described before do not consider this information. In principle, information about community structure could be useful for the fairness aspect of FAI.

Therefore, we consider seven community-aware centrality metrics (CACMs) that already exist in literature - Community based mediator [26], Community-Hub-Bridge[6], Comm Centrality [9], Participation Co-efficient [8], Modularity Vitality [17], Community K-shell[7], and Community based centrality [27]. As discussed later in the results, at $\lambda_c$, Community-Hub-Bridge stands out as particularly effective for FAI among these CACMs (figure A.5) in terms of performance. Thus, the focus is on Community-Hub-Bridge and the remainder of the community-aware centrality metrics are described in A.3.

Community-Hub-Bridge is designed to identify nodes which are hubs within their own community but also bridges to other communities[6]. In simple terms, hubs are nodes with high degree, while bridges are nodes that connect two or more poorly-connected components (communities) of the network. Furthermore, it introduces two key parameters: $\rho_{intra}$ and $\rho_{inter}$. The $\rho_{intra}$ is the fraction of its

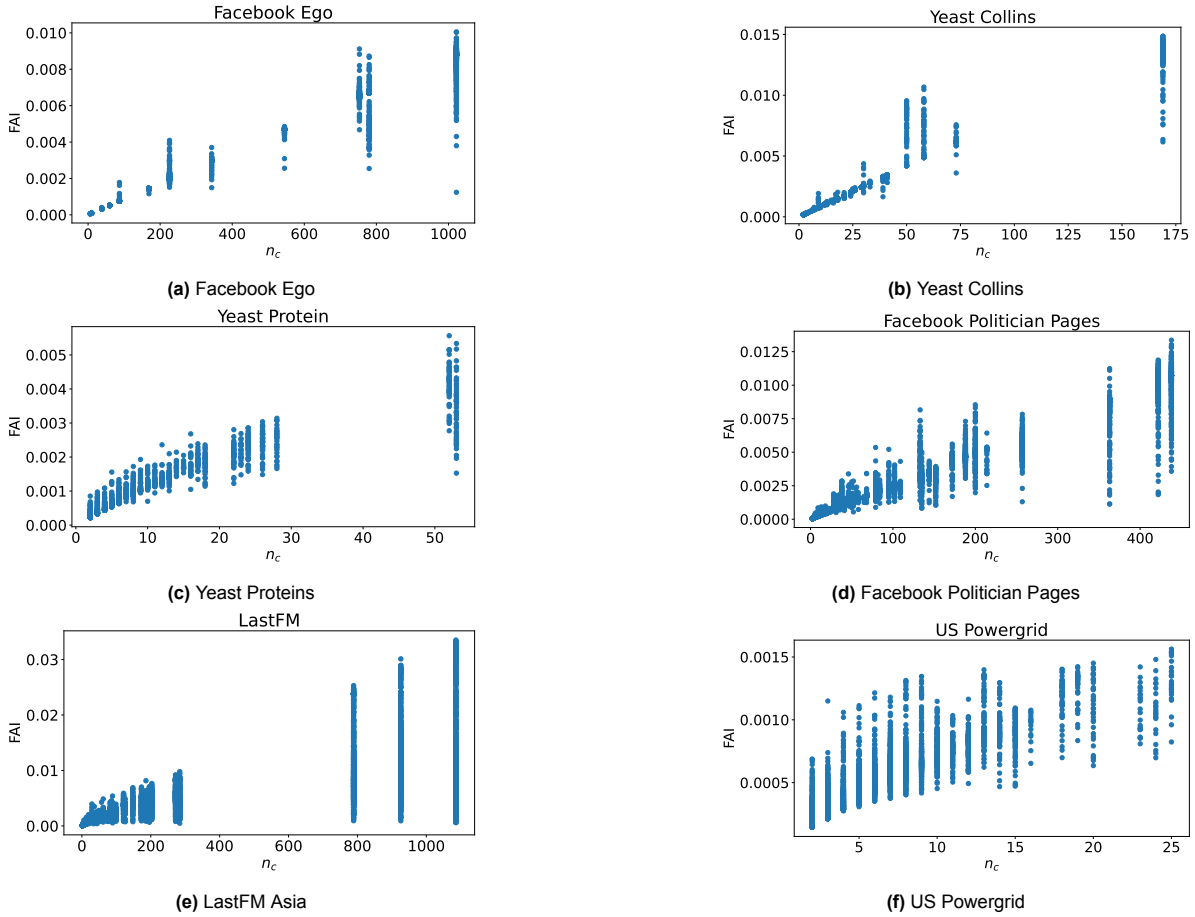
links to nodes within its own community to the total number of links. Similarly, $\rho_{inter}$ is the fraction of the links to nodes outside its own community. These parameters are essential in assessing the role of nodes within and between communities. A node's influence within its own community depends on $\rho_{intra}$ and on the size of its own community $n_{c_i}$. Its influence outside its community depends on $\rho_{inter}$ and the number of external communities it is connected to ($\beta_{NNC}$). Thus Community-Hub-Bridge($\chi$) of a node $i$ is defined as :

$$\chi(i) = n_{c_i} \times \rho_{intra} + \beta_{NNC} \times \rho_{inter}$$

## 4.5. New Metrics

Although existing CACMs incorporate community-related information, such as community size and link distribution within and outside community which could prove useful for measuring fairness. Consequently, while some of them are expected to outperform traditional metrics, they may not be the optimal choice for assessing FAI. Hence we propose two new classes of metrics designed specifically for FAI. Each class is designed using different approaches, each explained in 4.5.1 and 4.5.2.

The underlying assumption for these new metrics is that FAI can be explained using information from the local neighborhood of a seed node. This approach is motivated by the observation that at $\lambda = \lambda_c$, the infection rate is small and is mostly localized to the 2-hop neighborhood of the node (figures B.4, B.3). We consider two variants of the local neighborhood: one with direct neighbors (1-hop) and the other with neighbors and next nearest neighbors (2-hop). The seed node is also considered. As a result, each metric has two variants - $M^1$ and $M^2$.




**(a)** Facebook Ego

**(b)** Yeast Collins

**(c)** Yeast Proteins

**(d)** Facebook Politician Pages

**(e)** LastFM Asia

**(f)** US Powergrid

**Figure 4.2:** FAI of node v/s its own Community size

### 4.5.1. Class I

An interesting observation is made on the relationship between community size ($n_c$) and FAI(figure 4.2). In each real-world network, on average, the trend is that the FAI of a node is larger if its community size is larger. While it is expected for fairness to depend on the community size, the surprising aspect is that the correlation between $n_c$ and FAI is higher than the correlation between FAI and any other centrality metric as seen in figure 5.2. Thus, we can conclude that the size of the community that the seed node belongs to has a significant impact on its FAI.

In figure 4.2 we see that nodes with the same community size can have varying FAI. This is specially noticeable for nodes in large communities, where there is a large variation in FAI. This indicates that community size of the seed node alone cannot not accurately estimate FAI. The fairness of a node depends on the community size of the other nodes that it can possibly infect. This is evident from the formula of fairness 3.1 where we consider the fraction of number of infected nodes to community size i.e $p_c = \frac{i_c}{n_c}$. Thus, in addition to its own community size, we consider community size of its neighbors. The new metric can therefore be described as the average of community size of *all* its neighbors, including itself.

Let $L_1(i)$ be the set of nodes in the 1-hop neighborhood and $L_2(i)$ be the set of nodes in the 2-hop neighborhood, including the seed node $i$ in both cases. Let $n_{c_m}$ be the community size of node $m$. Thus, the respective community size-based metrics $M^1_{cs}$ and $M^2_{cs}$ of a given seed node $i$ can be defined as follows:

$$M^1_{cs}(i) = \frac{\sum_{m \epsilon L_1} \frac{n_{c_m}}{N}}{|L_1|} \tag{4.2}$$

$$M^2_{cs}(i) = \frac{\sum_{m \epsilon L_2} \frac{n_{c_m}}{N}}{|L_2|} \tag{4.3}$$

To select a node with high FAI at $\lambda_c$, the node itself should belong to a large community and its link distribution should be such that many of its links go to nodes with larger communities. A larger community tends to facilitate connections to more neighbors and is likely to improve prevalence. The fraction of infections in large communities tends to be small, in part due to the small number of infections at $\lambda_c$, which is likely to increase fairness.

### 4.5.2. Class II

By definition, FAI is consists of two components - prevalence and fairness. A different approach to designing a metric to estimate the FAI rank of the nodes would be to combine independent metrics that may indicate either the prevalence or fairness of a seed node.

For the prevalence of a seed node, we consider its local centrality. Chen [5] shows that $C_L$ of a seed node is strongly correlated with its prevalence in SIR infection spreading model for many graphs, outperforming degree and betweenness centralities. This observation holds true for the six real world networks considered (Figure 4.3).

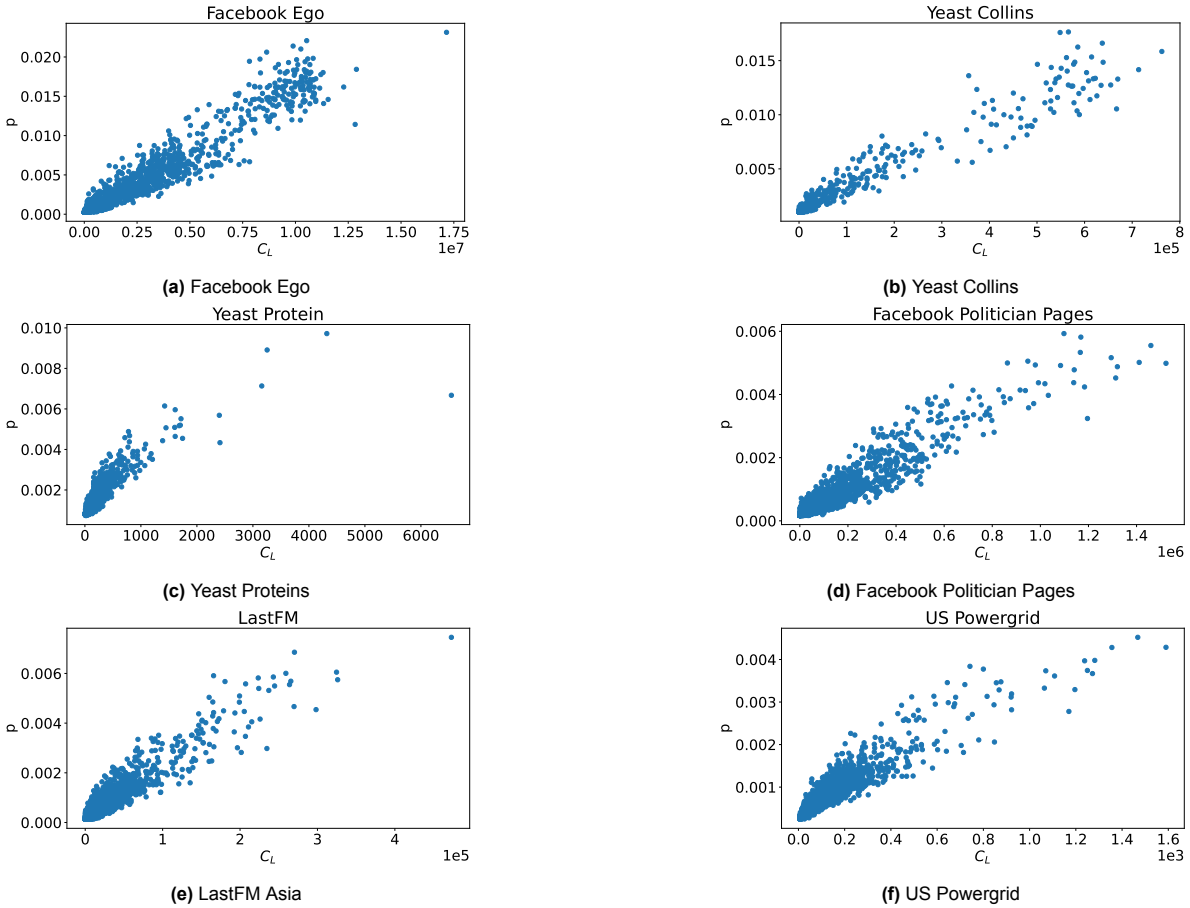
To estimate the fairness component, we use the following intuition. As stated earlier, we assume that at $\lambda_c$ the infection spread is usually restricted to nodes in one-hop or two-hop neighborhood of the seed node. We use another simplifying assumption - infection from a given seed node can spread to all the neighboring nodes. While this is improbable, this can help differentiate nodes in terms of FAI by considering all the possible communities the infection is likely to spread to. In other words, we assume that all the nodes in the neighborhood of the seed node get infected, and then compute the fairness using the fairness formula 3.1.

We then combine local centrality and fairness in the same fashion as the FAI formula (FAI = p/f).

Let $L_1(i)$ be the set of nodes in the 1-hop neighborhood and $L_2(i)$ be the set of nodes in the 2-hop neighborhood, including the seed node $i$ in both cases. Let $f^1$ and $f^2$ be the fairness when all nodes in the $L_1$ and $L_2$ neighborhood are infected respectively. Thus, the new metrics $M^1_{pf}$ and $M^2_{pf}$ may be defined as follows:

$$M^1_{pf}(i) = \frac{C_L(i)}{f_{X^1}} \tag{4.4}$$

$$M^2_{pf}(i) = \frac{C_L(i)}{f_{X^2}} \tag{4.5}$$




**(a)** Facebook Ego

**(b)** Yeast Collins

**(c)** Yeast Proteins

**(d)** Facebook Politician Pages

**(e)** LastFM Asia

**(f)** US Powergrid

**Figure 4.3:** Prevalence of node v/s its Local Centrality at $\lambda_c$. Each graph shows a strong correlation between Prevalence and Local Centrality.

## 4.6. Evaluation Methods

We generate two sets of ranks - one by ranking nodes based on FAI, and the other by ranking nodes based on a given metric. We employ each of the existing and newly proposed metrics for ranking nodes. To assess and compare the rankings produced by different centrality metrics, we utilize two methods: recognition rate and Kendall's rank correlation. Recognition rate facilitates the comparison of the top-ranked percentage of nodes 'k' between FAI and a given metric, while Kendall's rank correlation between FAI and a given metric offers an overall comparison of *all* the nodes. The value of these evaluation methods indicates how well a metric can rank nodes in terms of their FAI. The following sections describe these evaluation methods.

### 4.6.1. Recognition rate

It is defined as the number of nodes in common between the top '$k$' percent of nodes of two given rank lists. Let $r_1$ and $r_2$ be the set of highest ranked nodes, each set of size $int[k * n]$. Then recognition rate of top $f$ fraction of nodes $R(k)$ is defined as :

$$R(k) = \frac{r_1 \cap r_2}{|r_1|} \tag{4.6}$$

### 4.6.2. Kendall's Tau (B) - rank correlation

Kendall Tau B rank correlation($\tau$) [12] is used to measure dependence of the correlation between the objective rank generated on the basis of FAI (F), and the rank generated by a metric M. Since there may be ties in the ranking, especially with ranking based on the community size, Kendall's Tau-B variant is used to accommodate ties. Pearson's correlation is not used since the centrality metrics are not linearly correlated with FAI.

Thus, $\tau$ may be defined as:

$$\tau(F, M) = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}} \tag{4.7}$$

where

$$n_0 = n(n-1)/2$$

$$n_1 = \sum_i t_i(t_i - 1)/2$$

$$n_2 = \sum_j u_j(u_j - 1)/2$$

$n_c$ and $n_d$ are the number of concordant and discordant pairs respectively. $t_i$ and $u_j$ are the number of tied values in the i-th group of ties for the first quantity (F) and the j-th group of ties for the second quantity (M) respectively.

## 4.7. Experiment Setup

The experiments are implemented in python programming language. The graph properties and classic centralities are computed using the networkx [10] package. The SIR simulations are generated using EON fastSIR [11] module. A high level overview of the steps to compute the FAI and the recognition rate and correlation values for each node in the six networks are as follows.

First, some basic pre-processing steps are required. Some graphs contained self loops which were removed. For disconnected graphs, such as Yeast Collins and Yeast Proteins, the largest connected components were used. For the real world networks, louvain method [2] of community detection was used to extract communities for the networks.

Second, the epidemic threshold $\lambda_c$ is computed for each graph. The recovery rate is set to 1 ($\gamma = 1$). The effective transmission rate $\lambda$ is then equal to $\beta$. For a given value of $\lambda$, 100 SIR simulations were run for each seed node. For graphs with large number of nodes ($|N| > 2000$), 1000 nodes were randomly sampled to be the seed node. The average number of nodes that were infected at any given time during the epidemic spreading process were recorded. The variability measure $\Delta$ was computed for each $\lambda$. The value of $\lambda$ corresponding to the largest $\Delta$ value is $\lambda_c$. Using the Facebook Politician Pages and

the US Powergrid graphs as examples, the values of the variation $\Delta$ and infection rate values $\lambda$ are shown in figure 4.1.

Third, for each of the graphs, 100 runs of SIR simulations were performed for each node at the epidemic threshold. For each run, the fairness and prevalence were recorded at the end of the infection spreading process. The average of fairness and prevalence was computed over the 100 runs. It is important to average them over multiple runs since infection simulations are stochastic (some randomness is involved) and may result in very different fairness and prevalence values in each run. Then the FAI for each node was computed using this averaged prevalence and fairness for each node.

Fourth, the centrality Metrics, community aware metrics, $M_{cs}^1$, $M_{cs}^1$, $M_{pf}^1$ and $M_{pf}^2$ are computed for each node in the network. The nodes are ranked based on these metrics. The nodes are also ranked based on the FAI for each of the three values for transmission rate. The higher the FAI of the seed node, the better the rank. For a given transmission rate, the FAI rank is compared to rank based on each metric using rank correlation and recognition rate. These results are presented and discussed in chapter 5.

# 5

# Results and Discussion

## 5.1. Results

We present the findings of our research that relate to our two research questions. To reiterate, the two questions are - does FAI of a node depend on its network properties and how the performance of the metrics that capture such properties changes with increasing effective transmission rate. Before we dive into these results, we present an argument of why we study FAI at values around the epidemic threshold $\lambda_c$.

For each of the six networks, we compute the average $\mu_F$ and standard deviation $\sigma_F$ of FAI across all nodes at $\lambda_c$, $2\lambda_c$ and $3\lambda_c$. We find that the coefficient of variation or CV ($\frac{\sigma_F}{\mu_F}$%) of FAI increases as $\lambda$ decreases. Among the considered values of $\lambda$, CV is maximum at $\lambda_c$ for all the six real-world networks. This is summarized in table 5.1. Ranking nodes is more effective when the variation of FAI among the nodes is high. Thus we study the values of FAI around the epidemic threshold for values of $\lambda = \lambda_c, 2\lambda_c, 3\lambda_c$.

We compare the performance of the newly proposed metrics with the existing ones. This performance is measured in terms of the recognition rate and kendall's rank correlation. Since many metrics are involved, we pick two best performing metrics at $\lambda_c$ as the baseline, one from each of the two categories of existing metrics - simple centrality metrics and community-aware centrality metrics. The performance of the metrics belonging to each type is presented in 5.1.1 and 5.1.2 respectively. The reason for observed correlation between FAI and baseline metrics is presented in 5.2.2.

We present the correlation values between FAI and the baseline metrics ($C_L$, $\chi$), community size ($n_c$) and the newly proposed metrics in Table 5.2 and recognition rate plots at $\lambda_c$, $2\lambda_c$ and $3\lambda_c$ in Figures 5.1, A.6, and A.7 respectively. $n_c$ shows the highest correlation and one of the best recognition rates among all the metrics at $\lambda_c$, indicating it is an important property to consider to rank nodes in terms of FAI. However, this correlation decreases with increasing effective transmission rate.

Among the newly proposed metrics $M_{cs}^1$ , $M_{cs}^2$, $M_{pf}^1$ and $M_{pf}^2$, $M_{cs}^1$ performs the best across all six networks at $\lambda_c$. It also outperforms the baseline metrics. Its correlation with FAI decreases with increasing effective transmission rate.

These results also enable discussion of how the performance of the community size, baselines, and new metrics changes at higher levels of infection rate in 5.2. This helps to answer the second research question.

### 5.1.1. Performance of Simple Centrality Metrics

We consider the performance of degree, betweenness, closeness ($C_C$), eigenvector centrality ($C_E$) and local centrality $C_L$ in terms of recognition rate plots (figure A.4) and correlation values (Table A.1). It can be seen that degree centrality and betweenness centrality perform poorly for all graphs. $\tau(F, C_C)$ and $\tau(F, C_E)$ are the highest correlation values in Facebook Ego and LastFM networks respectively. For the remaining graphs, $\tau(F, C_L)$ is the highest correlation. Recognition rate of $C_C$ is the best for Facebook Ego, followed by $C_L$. For the remaining graphs, $C_E$ performs well for approximately top X=40% of the nodes, but for X > 40% it is outperformed by $C_L$. Since $C_L$ performs better than $C_E$ in multiple scenarios, it is taken as the baseline centrality metric.

| Network | $\lambda$ | $\mu_F$ | $\sigma_F$ | $\frac{\sigma_F}{\mu_F}$% |
|---|---|---|---|---|
| FacebookEgo | $\lambda_c$ | 0.00575 | 0.00250 | **43.41** |
| | $2\lambda_c$ | 0.00653 | 0.00270 | 41.33 |
| | $3\lambda_c$ | 0.00810 | 0.00307 | 37.93 |
| Yeast-Collins* | $\lambda_c$ | 0.00451 | 0.00455 | **100.98** |
| | $2\lambda_c$ | 0.00521 | 0.00500 | 96.00 |
| | $3\lambda_c$ | 0.00569 | 0.00503 | 88.49 |
| Yeast-proteins* | $\lambda_c$ | 0.00127 | 0.00099 | **77.74** |
| | $2\lambda_c$ | 0.00198 | 0.00122 | 61.50 |
| | $3\lambda_c$ | 0.00415 | 0.00193 | 46.59 |
| Facebook Politician Pages | $\lambda_c$ | 0.00394 | 0.00339 | **85.89** |
| | $2\lambda_c$ | 0.00578 | 0.00429 | 74.31 |
| | $3\lambda_c$ | 0.00933 | 0.00519 | 55.66 |
| Last-FM Asia | $\lambda_c$ | 0.00680 | 0.00862 | **126.64** |
| | $2\lambda_c$ | 0.00787 | 0.00575 | 73.08 |
| | $3\lambda_c$ | 0.01153 | 0.00529 | 45.84 |
| US Powergrid | $\lambda_c$ | 0.00050 | 0.00028 | **54.58** |
| | $2\lambda_c$ | 0.00082 | 0.00043 | 52.01 |
| | $3\lambda_c$ | 0.00152 | 0.00071 | 46.55 |

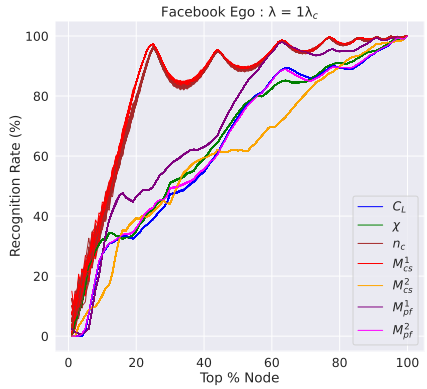
**Table 5.1:** Statistical measures of FAI at $\lambda_c$, $2\lambda_c$ and $3\lambda_c$ for the six real world networks. The Coefficient of variation is maximum at $\lambda_c$ for all the networks, and has been highlighted in **bold**.
$*$ means that the largest connected component is used since the graph is disconnected

### 5.1.2. Performance of Community-Aware Centrality Metrics

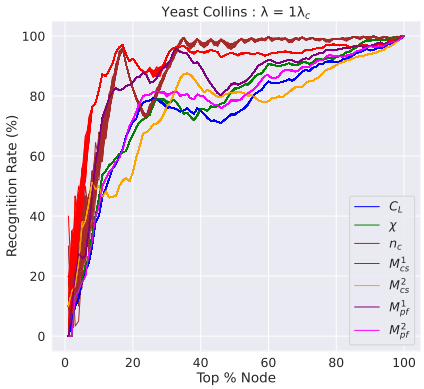
We consider the performance of seven existing community-aware centrality metrics. These are community-hub-bridge, participation co-efficient, community-based centrality, comm centrality, modularity vitality, community k-shell and community based mediator. The correlation between FAI and community-hub-bridge ($\chi$) is the highest among all other metrics as seen in Table A.2. In the case of the Facebook Ego network, $\chi$'s performance aligns closely with that of Community-based Centrality. However, in the other network instances, $\chi$ distinctly outperforms other metrics in terms of recognition rate A.5. As a result, $\chi$ is taken as the baseline community-aware centrality metric.

## 5.2. Discussion

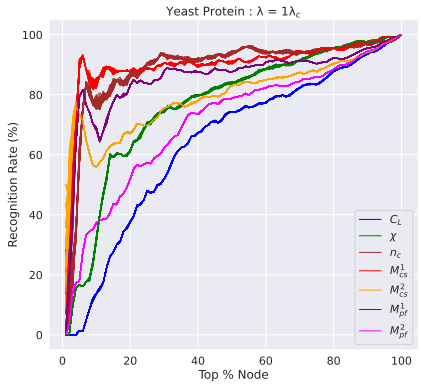
We observe that community size $n_c$ has the highest correlation with FAI and high recognition rate, among any other metric. The value of metric $\chi$ also depends partly on $n_c$. Thus it is important to analyze the impact of community size on FAI and explain how it varies with different infection levels in 5.2.1. We discuss the comparative performance of baselines ($C_L$ and $\chi$) at different infection levels and summarize the reasons for performance variations in 5.2.2 to address the second research question. We then explore the correlation and recognition rate results for the newly proposed metrics in 5.2.3, while highlighting the best-performing metric and its performance across different networks and infection levels. We also provide a concise summary of the major findings and their implications for FAI ranking and highlight the key takeaways from the discussion in 5.2.4.



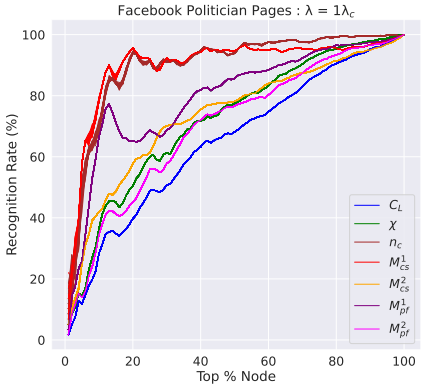

**(a)** Facebook Ego



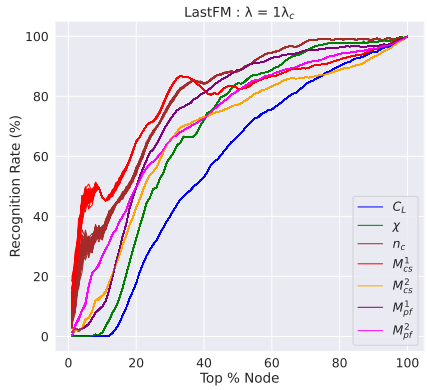

**(b)** Yeast Collins



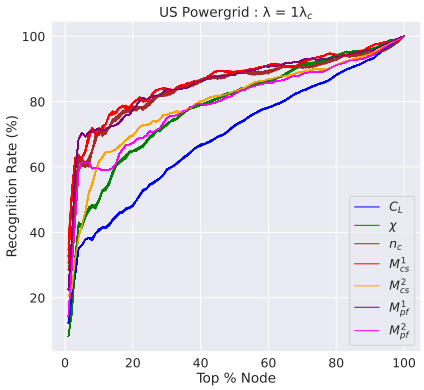

**(c)** Yeast Proteins




**(d)** Facebook Politician Pages




**(e)** LastFM Asia




**(f)** US Powergrid

**Figure 5.1:** Recognition Rate Plot at $\lambda_c$

| Network | $\lambda$ | $C_L$ | $\chi$ | $n_c(\downarrow)$ | $M_{cs}^1(\downarrow)$ | $M_{cs}^2$ | $M_{pf}^1$ | $M_{pf}^2$ |
|---|---|---|---|---|---|---|---|---|
| FacebookEgo | $\lambda_c$ | 0.41 | 0.43 | **0.87*** | **0.86** | 0.33 | 0.60 | 0.42 |
| | $2\lambda_c$ | 0.51 | 0.48 | **0.70*** | **0.69** | 0.22 | 0.39 | 0.49 |
| | $3\lambda_c$ | 0.46 | 0.40 | **0.58*** | **0.57** | 0.18 | 0.31 | 0.44 |
| Yeast-Collins* | $\lambda_c$ | 0.59 | 0.68 | **0.93*** | **0.87** | 0.55 | 0.77 | 0.66 |
| | $2\lambda_c$ | 0.62 | 0.67 | **0.87*** | **0.86** | 0.58 | 0.79 | 0.68 |
| | $3\lambda_c$ | 0.62 | 0.65 | **0.80** | **0.83*** | 0.60 | 0.78 | 0.69 |
| Yeast-proteins* | $\lambda_c$ | 0.41 | 0.71 | **0.86*** | **0.82** | 0.62 | 0.30 | 0.40 |
| | $2\lambda_c$ | 0.55 | 0.60 | **0.65** | **0.68*** | 0.64 | 0.25 | 0.51 |
| | $3\lambda_c$ | 0.58 | 0.52 | 0.55 | **0.60** | **0.61*** | 0.23 | 0.50 |
| Facebook Politician Pages | $\lambda_c$ | 0.40 | 0.59 | **0.91*** | **0.86** | 0.56 | 0.70 | 0.53 |
| | $2\lambda_c$ | 0.53 | 0.61 | **0.71** | **0.71** | 0.54 | **0.73*** | 0.63 |
| | $3\lambda_c$ | 0.59 | 0.61 | **0.62** | 0.63 | 0.51 | **0.72*** | 0.66 |
| Last-FM Asia | $\lambda_c$ | 0.29 | 0.61 | **0.79*** | **0.68** | 0.49 | **0.68** | 0.58 |
| | $2\lambda_c$ | 0.42 | 0.60 | **0.65** | 0.61 | 0.52 | **0.67*** | 0.63 |
| | $3\lambda_c$ | 0.44 | 0.59 | 0.62 | 0.58 | 0.51 | **0.67*** | **0.64** |
| US Powergrid | $\lambda_c$ | 0.45 | 0.68 | **0.77*** | **0.76** | 0.65 | 0.75 | 0.63 |
| | $2\lambda_c$ | 0.51 | 0.50 | 0.51 | 0.55 | 0.61 | **0.69*** | **0.66** |
| | $3\lambda_c$ | 0.42 | 0.34 | 0.34 | 0.38 | **0.46** | **0.50*** | **0.50*** |

**Table 5.2:** Correlation values of FAI at $\lambda_c$, $2\lambda_c$ and $3\lambda_c$ for the six real world networks.The highest (marked with *) and second-highest correlation values for each row has been highlighted in **bold**. Metrics that show a decreasing trend with increasing $\lambda$ are marked with ($\downarrow$).

### 5.2.1. Performance of Community size

At $\lambda_c$ the number of infections are limited as seen in figures B.3 and B.4. This indicates that infections are mostly confined to nodes within a two-hop range. In some simulations, the seed node recovers before infecting others. In context of small communities, the fraction of infection will be large compared to other bigger communities, leading to poor fairness. Thus at a community level, the fairness tends to be lesser for nodes of very small communities. Nodes in smaller communities also tend to have fewer connections to other nodes, which implies prevalence is likely to be smaller for such nodes. Therefore a high correlation and recognition rate is seen for community size at $\lambda_c$.

At higher effective transmission rates such as $2\lambda_c$ and $3\lambda_c$ this correlation is seen to decrease. As the infection rate increases, infection tends to spread to nodes in distant neighborhoods B.5. Considering only the size of the community the seed node belongs to as a metric does not suffice.

### 5.2.2. Performance of Baseline Metrics

Analysis of the correlation table and recognition rate plots(figure 5.1), reveals that, at $\lambda_c$, $\chi$ outperforms $C_L$. This suggests that incorporating community information improves estimation of ranks based on FAI. However at higher $\lambda$ values (from $3\lambda_c$ onwards) , $C_L$ surpasses $\chi$ in terms of correlation.

Two potential reasons for this shift can be hypothesized. First, higher $\lambda$ values lead to increased prevalence, which in turn increases the correlation with $C_L$. $C_L$ considers the information of the nodes in the 3-hop and 4-hop neighborhood. With a higher infection rate, more nodes become infected and the infection tends to spread further away the node. This leads to a stronger correlation with $C_L$.

The second reason is related to the correlation with community size, which decreases at higher level of $\lambda$. $\chi$ relies on community size of the seed node, and considers the count of the communities

reachable in one hop. This limited scope affects its ability to rank nodes by FAI. It is possible that $\chi$'s performance can be improved by considering additional number of hops based on the effective transmission rate, but it falls outside the scope of this project.

### 5.2.3. Performance of New Metrics

The performance of the new metrics are described in two parts - the correlation table and recognition rate.

#### Correlation

At $\lambda_c$, $n_c$ is seen to have the highest correlation, followed by $M^1_{cs}$ outperforming the rest. This is because at $\lambda_c$, in most simulation runs, the seed node recovers before it can infect other nodes or infects few other nodes in its 1-hop neighborhood. This can be observed in the examples shown in figures B.1 and B.3. It rarely infects a large number of nodes. This is also why the 1-hop network metrics outperform their respective 2-hop counterparts. If the infection rate were to increase, the correlation between FAI and $M^1$ metrics should reduce while increasing for $M^2$. This trend is observed at higher levels of effective transmission rate in table 5.2.

The performance of the 2-hop metrics is optimal when infection predominantly occurs within the 2-hop neighborhood. However, at $\lambda_c$, using the 2-hop neighborhood information results in less reliable estimates, given that many seed nodes (especially with smaller degree) tend to not infect their two-hop neighbors. The performance of 2-hop metrics diminishes at higher infection rates, as infection spreads to the 3-hop neighborhood and beyond. Nevertheless, the specific correlation trends vary across networks, as the spread of infection depends on factors beyond the effective transmission rate. For example, in Facebook Ego and US Powergrid, $M^2_{pf}$ is minimum at $\lambda_c$, increases at $2\lambda_c$ and decreases again at $3\lambda_c$, while in Yeast Collins and Facebook Political Pages, the correlation continues to increase.

Between the two classes of metrics, it is clear that $M^1_{cs}$ shows better performance than $M^1_{pf}$ at $\lambda_c$. The choice between $M^1_{pf}$ and $M^1_{cs}$ becomes more complex at higher infection rates, as their performance varies among different graphs. This is attributed, in part, to the stochastic nature of the infection spreading process. At higher infection rates, more nodes become susceptible to infection, and it is difficult to estimate which of these nodes become infected. It is not necessary that *all* nodes in a lower hop neighborhood will be infected before nodes in higher hop neighborhood. In such cases, considering larger neighborhoods is insufficient, and an alternative approach for metric design is recommended. Thus, the presented metrics are stable only at the epidemic threshold.

#### Recognition Rate Plots

From the recognition rate plots at $\lambda_c$ (figure 5.1) it is observed that $M^1_{cs}$ (red plot) is one of the best metrics at identifying top 20% of the nodes with highest FAI. In some graphs, it is outperformed by $n_c$, but the performance is still comparable. $M^1_{pf}$ performs slightly worse than the top 2 metrics except in Facebook Ego, where it is significantly worse for top 20% to top 45% of the nodes.

At $2\lambda_c$, $M^1_{pf}$ outperforms the rest in Yeast Proteins and US Powergrid, and performs comparably with $M^1_{cs}$ and $n_c$ in almost all the remaining networks. However, for Facebook Ego, it overtakes $M^1_{cs}$ and $n_c$ only after top 40% of the nodes. At $3\lambda_c$, $M^1_{pf}$ is the best metric for all networks except in Facebook Ego.

It is important to highlight that the Facebook Ego network exhibits somewhat peculiar behavior compared to the other networks. We see that correlation of $M^2_{cs}$ is low compared to other graphs, closeness centrality performing better than $C_L$, etc. This variation may be attributed to the extremely low mixing parameter ($\mu$=0.02), very high average degree ($< k >$=43.69) or the small number of communities, but it is difficult to pinpoint the exact reasons.

### 5.2.4. Overall Summary

To summarize the discussion, $M^1_{cs}$ demonstrates good performance in terms of both correlation and recognition rate at $\lambda_c$ across all networks. This suggests that, for small levels of infection, a good performance at ranking nodes in terms of FAI can be achieved by considering local community information of the seed node. Consequently, we can infer a dependency of network properties on FAI.

The FAI of a node changes with higher levels of infection, potentially affecting the performance of the metrics used to rank the node. While the newly proposed metrics perform well at $\lambda_c$, their

performance diminishes at higher levels of $\lambda$. Estimating fairness becomes challenging as it becomes difficult to predict which nodes in specific communities may be infected. Thus, this addresses our second research question, elucidating the performance variation of metrics with infection rate.

We also conclude that FAI of a node is dependent on various factors such as properties of the node and the network topology. The observed variations in performance across different networks are anticipated, considering their diverse global network properties, such as average degree and the number of communities. However, definitively pinpointing the specific factors influencing performance and their precise impact poses a challenging task.

# 6

# Conclusion

In summary, our contributions include introducing novel fairness-aware-influence (FAI) metrics to identify nodes with high prevalence and high fairness. At low levels of infection it is sufficient to consider information from the local neighborhood of the seed node to rank it in terms of FAI. We demonstrate that incorporating relevant community information, specially the size of the community to which the seed node belongs, enhances metric performance in ranking nodes. The newly proposed metric $M_{cs}^{1}$ exceeds the performance of all other existing centrality- and community aware metrics at the epidemic threshold. Its performance correlation diminishes with higher transmission rates, suggesting the need to consider larger neighborhoods in such cases. However, there is another problem at high transmission rates - a large number of nodes are susceptible to infection, but it is difficult to estimate which nodes will get infected. As future work, we recommend exploring new metrics that offer stable performance across a range of transmission rates.

Our current work explores the FAI of each individual node. As an extension, it would be interesting to study multi-spreader FAI problem at the epidemic threshold, where a fraction of nodes are considered as seed nodes and each set of seed nodes must be ranked in terms of FAI. Perhaps the same metrics proposed by us can be calculated for each seed node and then aggregated in some manner to rank a set of seed nodes.

One limitation is our inability to explain the differing performance of the Facebook Ego graph compared to other graphs. This variation may be due to the extremely low mixing parameter ($\mu$=0.02) or high average degree ($< k >$=43.69). We suggest conducting an analysis of a wide range of global network features to explain these differences among graphs.

# References


[1] Junaid Ali et al. “On the Fairness of Time-Critical Influence Maximization in Social Networks”. In: *IEEE Transactions on Knowledge and Data Engineering* 35.3 (2023), pp. 2875–2886. DOI: 10.1109/TKDE.2021.3120561.

[2] Vincent D Blondel et al. “Fast unfolding of communities in large networks”. In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[3] Doina Bucur. “Top influencers can be identified universally by combining classical centralities”. In: *CoRR* abs/2006.07657 (2020). arXiv: 2006.07657. URL: https://arxiv.org/abs/2006.07657.

[4] Deepayan Chakrabarti et al. “Epidemic Thresholds in Real Networks”. In: *ACM Trans. Inf. Syst. Secur.* 10.4 (Jan. 2008). ISSN: 1094-9224. DOI: 10.1145/1284680.1284681. URL: https://doi.org/10.1145/1284680.1284681.

[5] Duanbing Chen et al. “Identifying influential nodes in complex networks”. In: *Physica A: Statistical Mechanics and its Applications* 391.4 (2012), pp. 1777–1787. ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2011.09.017. URL: https://www.sciencedirect.com/science/article/pii/S0378437111007333.

[6] Zakariya Ghalmane, Mohammed El Hassouni, and Hocine Cherifi. “Immunization of networks with non-overlapping community structure”. In: *Social Network Analysis and Mining* 9 (2019), pp. 1–22.

[7] Kai Gong and Li Kang. “A New K-Shell Decomposition Method for Identifying Influential Spreaders of Epidemics on Community Networks”. In: *Journal of Systems Science and Information* 6.4 (Sept. 2018), pp. 366–375. DOI: 10.21078/jssi-2018-366-10. URL: https://doi.org/10.21078%2Fjssi-2018-366-10.

[8] Roger Guimera and Luís A Nunes Amaral. “Functional cartography of complex metabolic networks”. In: *nature* 433.7028 (2005), pp. 895–900.

[9] Naveen Gupta, Anurag Singh, and Hocine Cherifi. “Centrality measures for networks with community structure”. In: *Physica A: Statistical Mechanics and its Applications* 452 (2016), pp. 46–59.

[10] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. “Exploring Network Structure, Dynamics, and Function using NetworkX”. In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, 2008, pp. 11–15.

[11] Z Istvan et al. *Mathematics of Epidemics on Networks: From Exact to Approximate Models*. Springer, 2019.

[12] M. G. Kendall. “The Treatment of Ties in Ranking Problems”. In: *Biometrika* 33.3 (1945), pp. 239–251. ISSN: 00063444. URL: http://www.jstor.org/stable/2332303 (visited on 09/26/2023).

[13] Jérôme Kunegis. “Handbook of Network Analysis [KONECT - the Koblenz Network Collection]”. In: *CoRR* abs/1402.5500 (2014). arXiv: 1402.5500. URL: http://arxiv.org/abs/1402.5500.

[14] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. “Benchmark graphs for testing community detection algorithms”. In: *Physical review E* 78.4 (2008), p. 046110.

[15] Jure Leskovec and Andrej Krevl. *SNAP Datasets: Stanford Large Network Dataset Collection*. http://snap.stanford.edu/data. June 2014.

[16] Jianxin Li et al. “Community-diversified influence maximization in social networks”. In: *Information Systems* 92 (2020), p. 101522. ISSN: 0306-4379. DOI: https://doi.org/10.1016/j.is.2020.101522. URL: https://www.sciencedirect.com/science/article/pii/S0306437920300326.

[17] Thomas Magelinski, Mihovil Bartulovic, and Kathleen M. Carley. “Measuring Node Contribution to Community Structure With Modularity Vitality”. In: *IEEE Transactions on Network Science and Engineering* 8.1 (2021), pp. 707–723. DOI: 10.1109/TNSE.2020.3049068.

[18] M. E. J. Newman. “The Structure and Function of Complex Networks”. In: *SIAM Review* 45.2 (2003), pp. 167–256. DOI: 10.1137/S003614450342480. eprint: https://doi.org/10.1137/S003614450342480. URL: https://doi.org/10.1137/S003614450342480.

[19] Kazuya Okamoto, Wei Chen, and Xiang-Yang Li. “Ranking of closeness centrality for large-scale social networks”. In: *Lecture Notes in Computer Science* 5059 (2008), pp. 186–195.

[20] Tiago P. Peixoto. *The Netzschleuder network catalogue and repository*. URL: https://networks.skewed.de/net/collins_yeast.

[21] Stephany Rajeh et al. “Comparative evaluation of community-aware centrality measures”. In: *Quality & Quantity* 57.2 (2023), pp. 1273–1302.

[22] Ryan Rossi and Nesreen Ahmed. “The Network Data Repository with Interactive Graph Analytics and Visualization”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 29.1 (Mar. 2015). DOI: 10.1609/aaai.v29i1.9277. URL: https://ojs.aaai.org/index.php/AAAI/article/view/9277.

[23] Britta Ruhnau. “Eigenvector-centrality — a node-centrality?” In: *Social Networks* 22.4 (2000), pp. 357–365. ISSN: 0378-8733. DOI: https://doi.org/10.1016/S0378-8733(00)00031-9. URL: https://www.sciencedirect.com/science/article/pii/S0378873300000319.

[24] Fangshuang Tang et al. “Diversified social influence maximization”. In: *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. 2014, pp. 455–459. DOI: 10.1109/ASONAM.2014.6921625.

[25] Alan Tsang et al. “Group-Fairness in Influence Maximization”. In: *CoRR* abs/1903.00967 (2019). arXiv: 1903.00967. URL: http://arxiv.org/abs/1903.00967.

[26] Muluneh Mekonnen Tulu, Ronghui Hou, and Talha Younas. “Identifying Influential Nodes Based on Community Structure to Speed up the Dissemination of Information in Complex Network”. In: *IEEE Access* 6 (2018), pp. 7390–7401. DOI: 10.1109/ACCESS.2018.2794324.

[27] Zhiying Zhao et al. “A community-based approach to identifying influential spreaders”. In: *Entropy* 17.4 (2015), pp. 2228–2252.

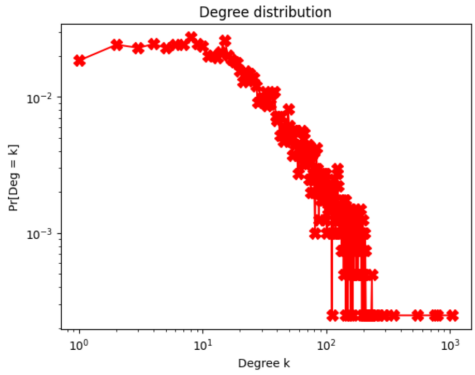
# A

# Appendix

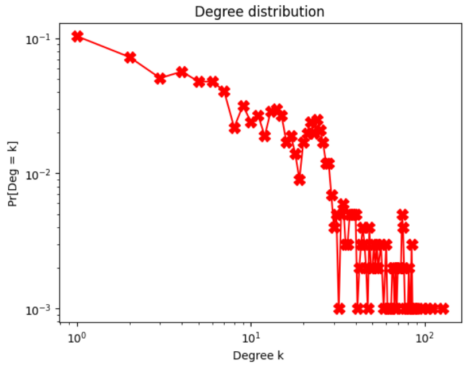
## A.1. Network Information

This section presents the degree distribution and community size distribution of the six real world networks. The degree distribution is plotted in a logarithmic scale, typically to see if these networks are scale-free networks or not. Scale-free networks show a degree distribution that follows a power law. Thus, the plot should be close to a straight line. It is difficult to conclude that the six real world networks are scale-free.

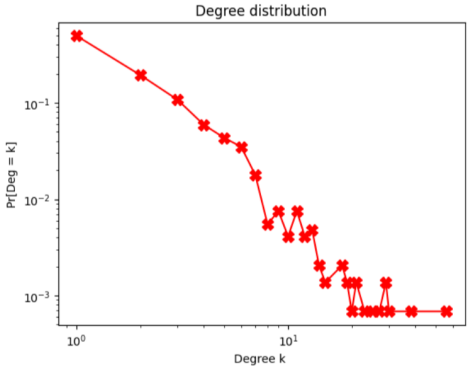
The community distribution plots show that these networks have a heterogeneous distribution of community size. There are many small communities and few large communities. Facebook Ego and LastFM Asia have one of the largest communities among all other networks, of size approximately 1000.



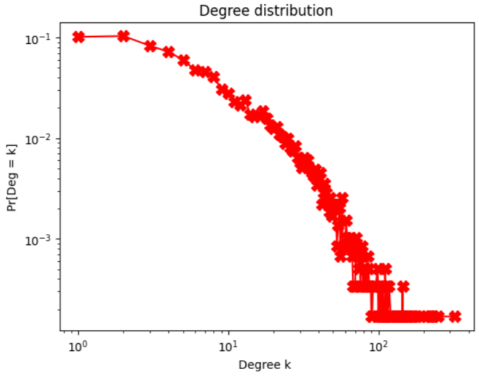

**(a)** Facebook Ego



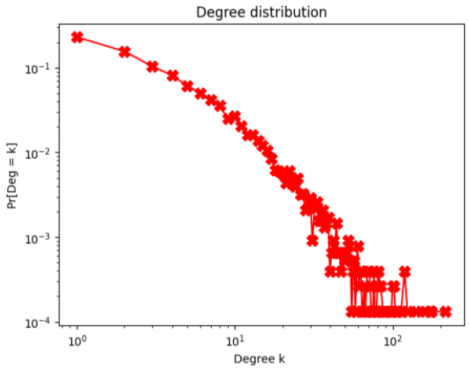

**(b)** Yeast Collins



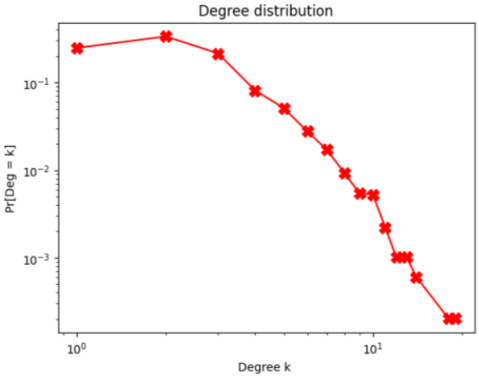

**(c)** Yeast Proteins




**(d)** Facebook Politician Pages

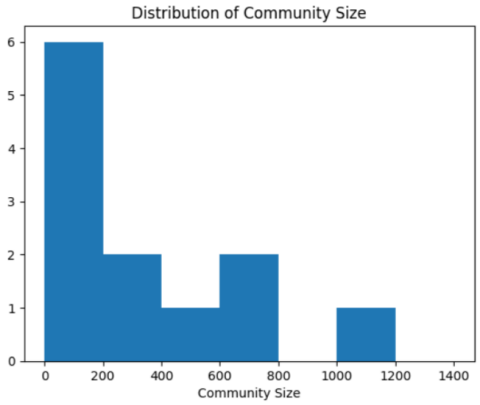



**(e)** LastFM Asia

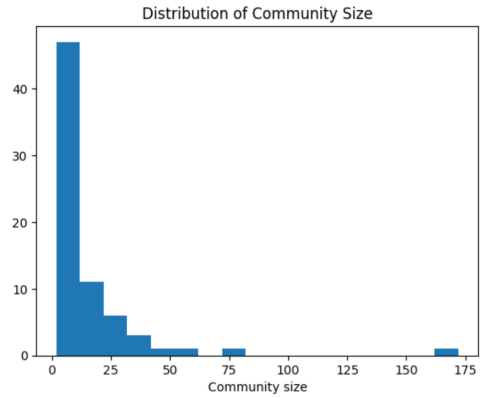



**(f)** US Powergrid

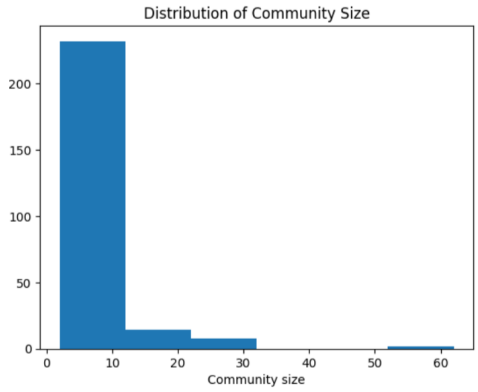
**Figure A.1:** Degree Distribution for the real world networks



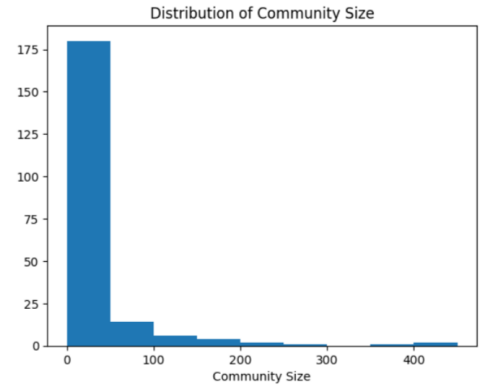

**(a)** Facebook Ego



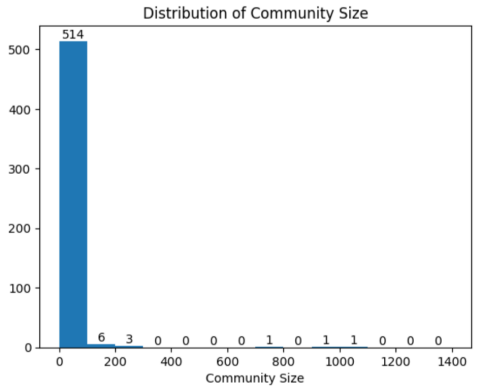

**(b)** Yeast Collins



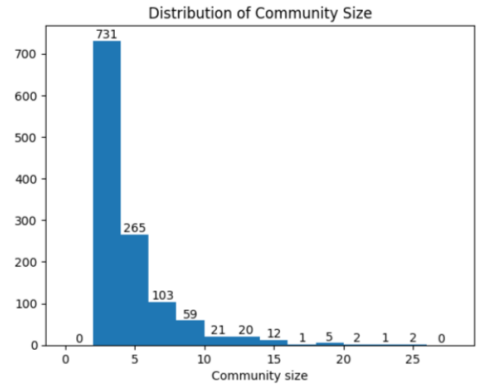

**(c)** Yeast Proteins




**(d)** Facebook Politician Pages




**(e)** LastFM Asia




**(f)** US Powergrid

**Figure A.2:** F-SD v/s Prevalence for real world networks

## A.2. Relationship between Fairness and Prevalence

It is interesting to see the relationship between fairness and prevalence at $\lambda_c$. There are some nodes which have low prevalence and poor fairness. A general trend is seen where SD of variation of seed nodes increases with increase in prevalence.

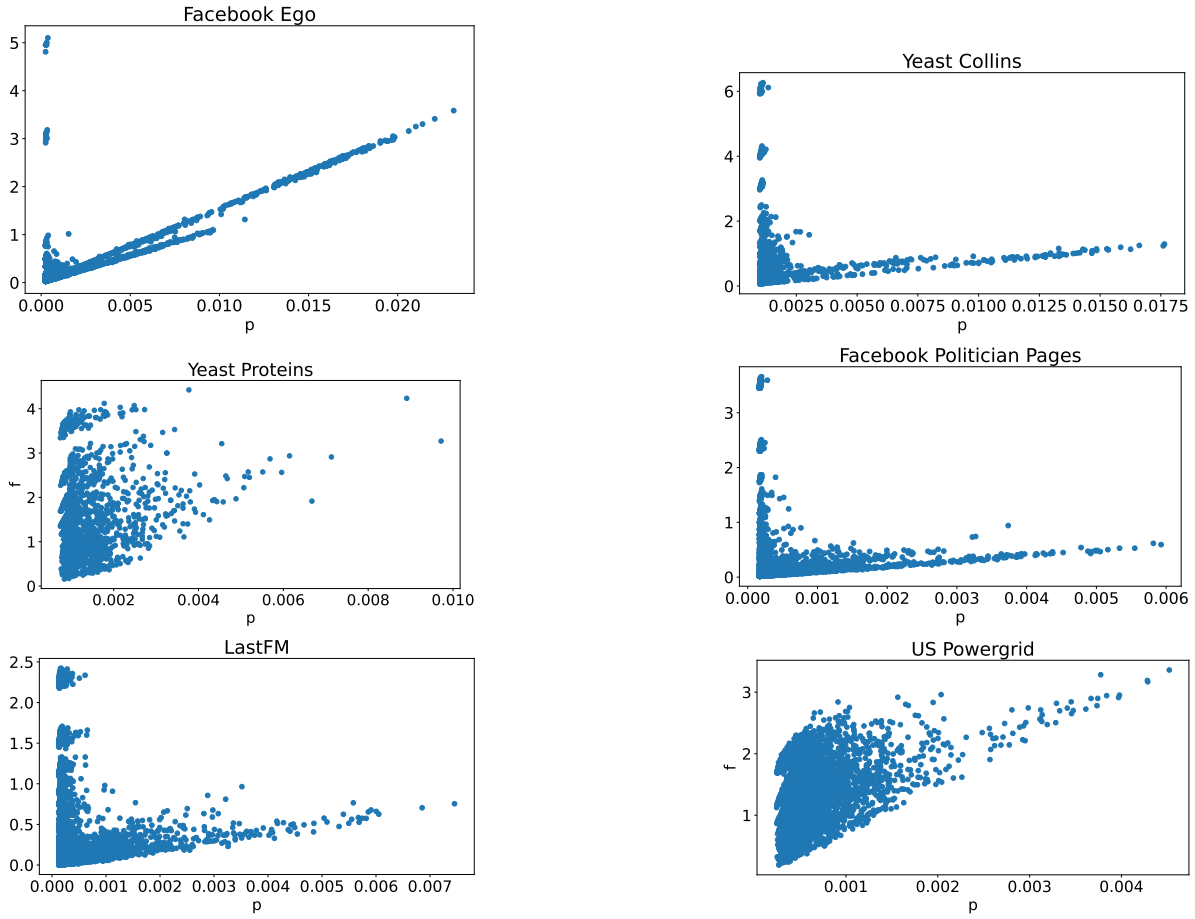



**Figure A.3:** F-SD v/s Prevalence for real world networks at $\lambda_c$

## A.3. Relation between existing Centrality metrics and FAI

The existing community aware centrality metrics are briefly decribed as follows.

### A.3.1. Community-based centrality

Community-based Centrality($C_1$) [27] weights the intra-community and inter-community links with the respective size of the community.Let $k_{ij}$ be the number of links from node $i$ to community $j$, then Community-based Centrality is defined as:

$$C_1(i) = \sum_{j \epsilon c} k_{ij} \times \frac{n_{c_j}}{N}$$

### A.3.2. Comm Centrality

Comm Centrality($C_2$) [9] uses information only at the community level. Let $\mu_{c_i}$ be the fraction of links outside the community of $c_i$ to the total number of links of all the nodes in the community $c_i$. It is defined as follows:

$$C_2(i) = (1 + \mu_{c_i}) \times \left( \frac{k_i^{intra}}{max_j(k_{j \epsilon c_i}^{intra})} \times R \right) + (1 - \mu_{c_i}) \times \left( \frac{k_i^{inter}}{max_j(k_{j \epsilon c_i}^{intre})} \times R \right)$$

where

$$\mu_{c_i} = \frac{\sum_{j \epsilon c_i} k_j^{inter}}{n_{c_i}}$$

and R is an integer chosen such that both intra and inter values are in the same range.

### A.3.3. Participation Co-efficient

The Participation Co-efficient($C_3$) is designed to differentiate nodes based on their role in the network [8]. The PC of a given node $i$ is defined as follows.

$$C_3(i) = 1 - \sum_{q=1}^{C} \left( \frac{k_{i,c_q}}{k_i^{tot}} \right)^2$$

The value is close to 1 if the edges of the node are uniformly distributed among all the communities, and 0 if all its edges are within its own community.

### A.3.4. Community-based mediator

Community-based mediator ($C_4$) considers the entropy of the inter- and intra-community links of a given node. An assumption is made that the inter-community links from a node to different communities are considered separately though not clearly shown in the equation (11) in the original paper[26]. Let $\rho_i$ denote the ratio of the links of $i$ to nodes within its own community $c_i$ to the total number of links of $i$. Let $\rho_{ij}$ be the ratio of links from node $i$ to community $j$ to the total number of links of i. Note that $j$ is any community in the network which is not the community of $i$. Thus the overall entropy $H_i$ of the various links of a node $i$ belonging to community $c_i$ is given as

$$H_i = [-\rho_i^{intra} log(\rho_i^{intra})] + [- \sum_{j \epsilon c \setminus c_i} \rho_{ij}^{inter} log(\rho_{ij}^{inter})]$$

Then, Community-based mediator of node $i$ is calculated as follows:

$$C_4(i) = H_i \times \frac{k_i^{tot}}{\sum_{i=1}^{N} k_i^{tot}}$$

The paper [26] claim that nodes selected by $C_4$ are key nodes to spread information in the network quickly and outperforms $C_1$ A.3.1.

### A.3.5. K-shell community

K-shell with community or $C_5$ [7] partitions the network into two - one with node and its own community and the other with the node and the remaining network. The original K-shell does not distinguish between strong and weak ties.

$$C_5(i) = \delta \times \alpha^{intra}(i) + (1 - \delta) \times \alpha^{inter}(i)$$

### A.3.6. Modularity Vitality

Rather than considering inter-and intra-community links, Modularity vitality ($C_6$) [17] studies the contribution of the node to the structure of the network using modularity. As mentioned earlier, modularity is a common metric to measure the goodness of community structure in a network. It may used to differentiate between hub and bridge nodes.

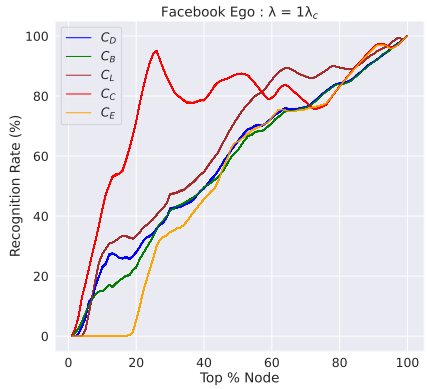
$$C_6(i) = Q(G) - Q(G - i)$$

| **Network** | $\lambda$ | $C_D$ | $C_B$ | $C_L$ | $C_C$ | $C_E$ |
|---|---|---|---|---|---|---|
| FacebookEgo | $\lambda_c$ | 0.23 | 0.17 | 0.41 | 0.55 | 0.27 |
| Yeast-Collins* | $\lambda_c$ | 0.40 | 0.12 | 0.59 | 0.45 | 0.50 |
| Yeast-proteins* | $\lambda_c$ | 0.06 | 0.06 | 0.41 | 0.38 | 0.41 |
| Facebook Politician Pages | $\lambda_c$ | 0.20 | 0.02 | 0.40 | 0.31 | 0.39 |
| Last-FM Asia | $\lambda_c$ | 0.07 | -0.04 | 0.29 | 0.26 | 0.34 |
| US Powergrid | $\lambda_c$ | 0.23 | 0.13 | 0.45 | 0.20 | 0.40 |

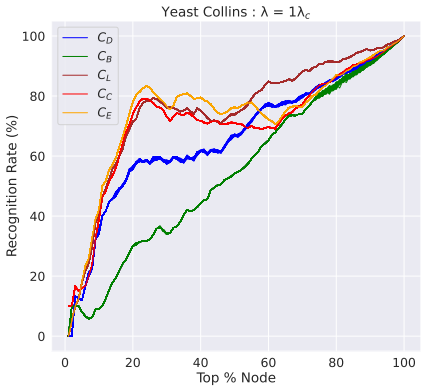
**Table A.1:** Correlation values of FAI with classic centrality metrics (degree $C_D$, betweenness $C_B$, closeness $C_C$, eigenvector $C_E$) and local centrality $C_L$ at $\lambda_c$ for the six real world networks. $C_L$ is seen to have the highest correlation values for all networks except LastFM Asia and Facebook Ego, and is chosen as the baseline for metrics without community information.

| **Network** | $\lambda$ | $\chi$ | $C_1$ | $C_2$ | | $C_3$ $C_4$ | $C_5$ | $C_6$ |
|---|---|---|---|---|---|---|---|---|
| FacebookEgo | $\lambda_c$ | **0.43** | 0.42 | -0.12 | -0.14 | -0.07 | 0.25 | -0.34 |
| Yeast-Collins* | $\lambda_c$ | **0.68** | 0.64 | 0.32 | 0.02 | 0.16 | 0.40 | -0.12 |
| Yeast-proteins* | $\lambda_c$ | **0.71** | 0.52 | -0.53 | -0.00 | 0.04 | 0.07 | -0.22 |
| Facebook Politician Pages | $\lambda_c$ | **0.59** | 0.52 | -0.40 | -0.13 | -0.03 | 0.23 | -0.05 |
| Last-FM Asia | $\lambda_c$ | **0.61** | 0.45 | -0.60 | -0.28 | -0.15 | 0.09 | -0.01 |
| US Powergrid | $\lambda_c$ | **0.68** | 0.56 | -0.35 | -0.06 | 0.10 | 0.18 | 0.00 |

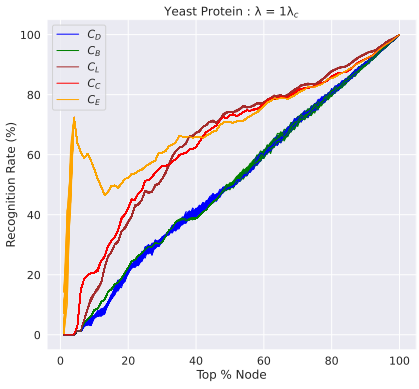
**Table A.2:** Correlation values of FAI with Community Aware Metrics at $\lambda_c$ for the six real world networks. $\chi$ is seen to have the highest correlation values for all networks. Thus, it is chosen as one of the baseline metrics.



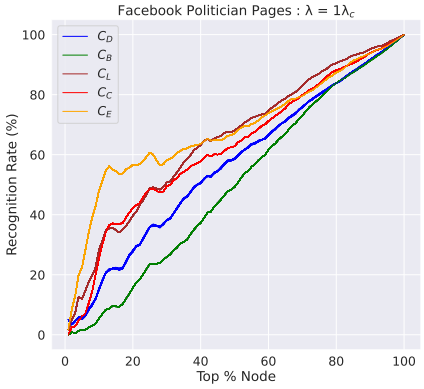

**(a)** Facebook Ego



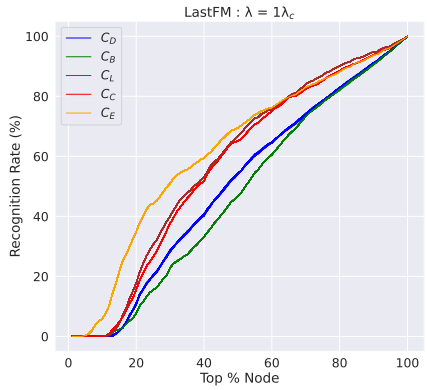

**(b)** Yeast Collins



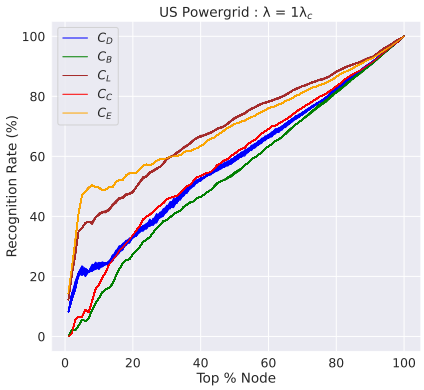

**(c)** Yeast Proteins




**(d)** Facebook Politician Pages

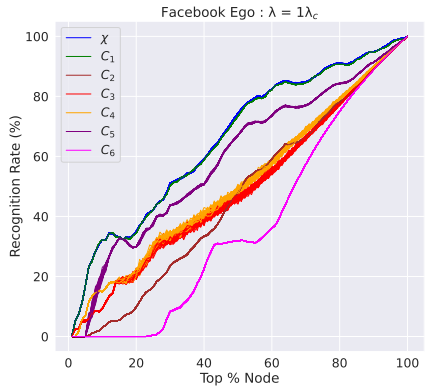



**(e)** LastFM Asia

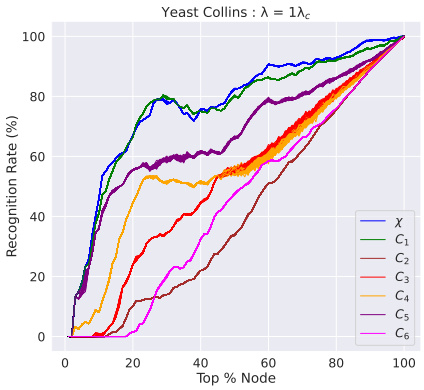



**(f)** US Powergrid

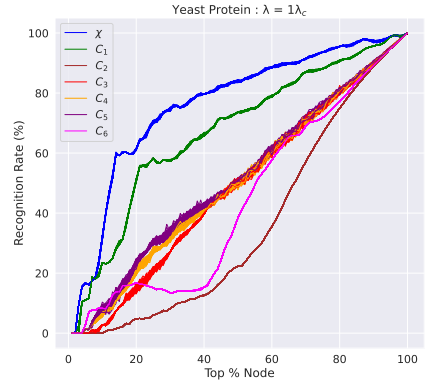
**Figure A.4:** Recognition rate plots for classic centrality Metrics and $C_L$ at $\lambda_c$



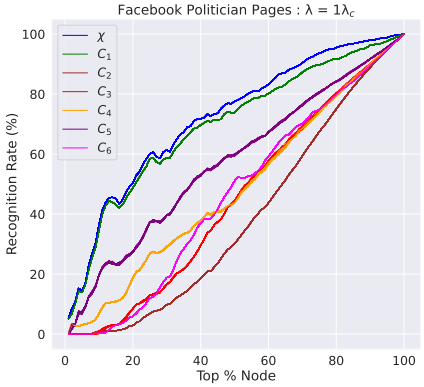

**(a)** Facebook Ego



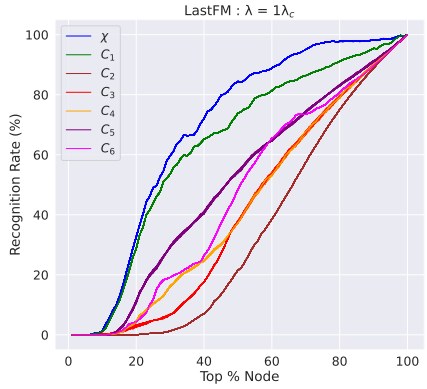

**(b)** Yeast Collins



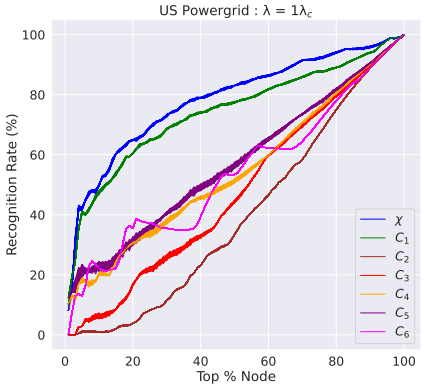

**(c)** Yeast Proteins




**(d)** Facebook Politician Pages




**(e)** LastFM Asia

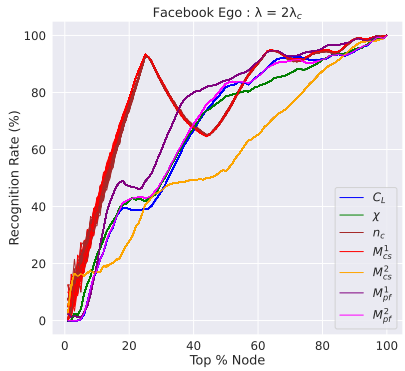



**(f)** US Powergrid

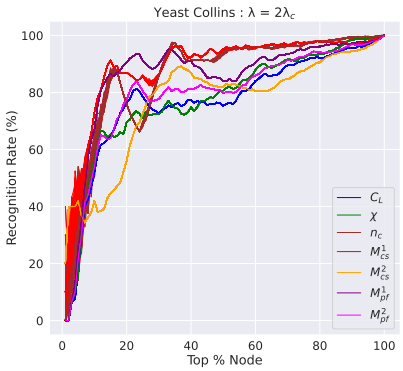
**Figure A.5:** Recognition Rate Plots for Community Metrics at $\lambda_c$. Community-Hub-Bridge ($\chi$) outperforms the others for each of the networks.

## A.4. FAI at higher levels of infection

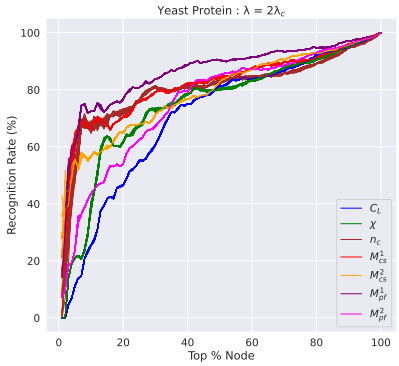
This section contains the recognition rate plots for the baselines and newly proposed metrics at $2\lambda_c$ and $3\lambda_c$.



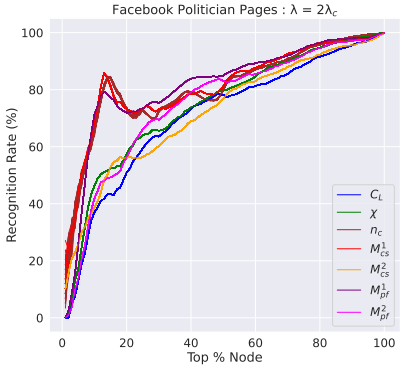

**(a)** Facebook Ego



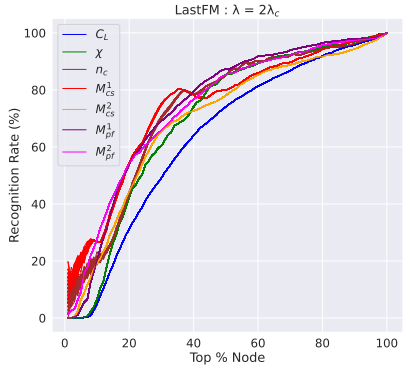

**(b)** Yeast Collins



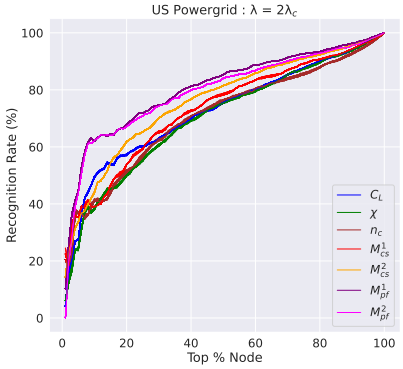

**(c)** Yeast Proteins




**(d)** Facebook Politician Pages

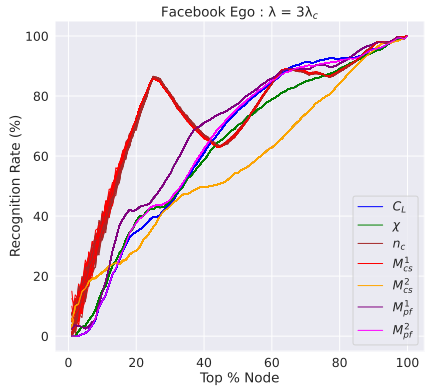



**(e)** LastFM Asia

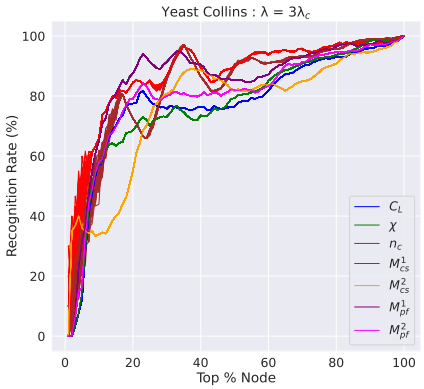



**(f)** US Powergrid

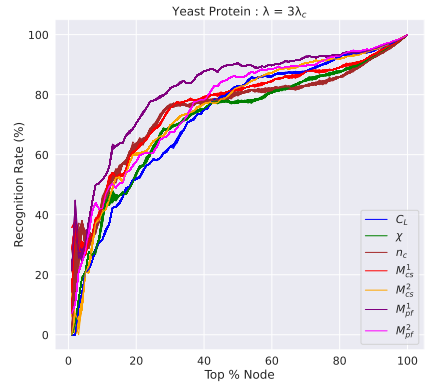
**Figure A.6:** Recognition Rate Plot for the baselines and newly proposed metrics at $2\lambda_c$



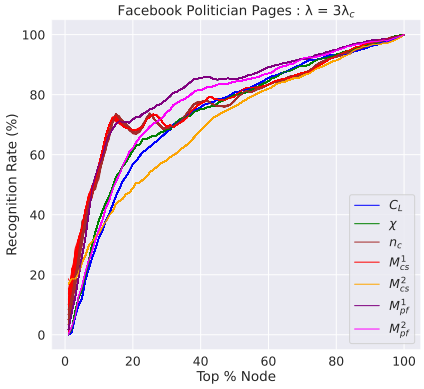

**(a)** Facebook Ego



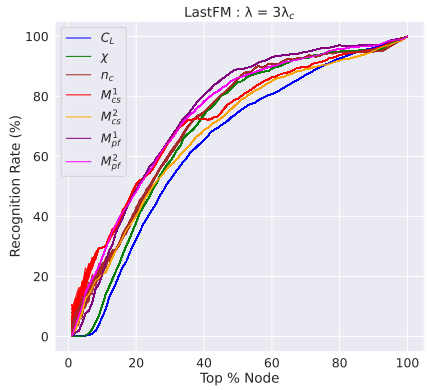

**(b)** Yeast Collins



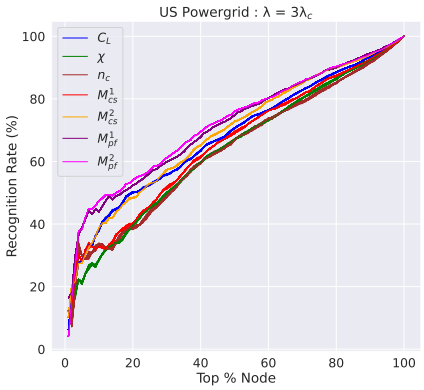

**(c)** Yeast Proteins




**(d)** Facebook Politician Pages




**(e)** LastFM Asia




**(f)** US Powergrid

**Figure A.7:** Recognition Rate Plot for the baselines and newly proposed metrics at $3\lambda_c$

# B

# SIR Simulations

To understand infection spread given a seed node, two examples of SIR simulations are presented in figuresB.1 and B.2. Two nodes with drastically different degrees from the Yeast Protein network are chosen. Node 1121 has a degree of 5 while node 224 has the highest degree of 56 in Yeast Proteins. It can be seen from the simulations that, on average, the number of infections is slightly more for node 224 due to the large number of neighbors. Another observation is that infections at $\lambda_c$ are mostly limited to one-hop neighborhood for both nodes. For nodes with larger degree, it is possible for the infection to spread to 2-hop neighbors as seen in B.4. This forms the basis on which the new metrics designed consider the 2-hop ego networks of a given seed node to rank it in terms of FAI.

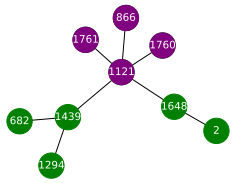



**Figure B.1:** A 2-hop ego network of node 1121 of the Yeast Protein network. Nodes with the same color as 1121 belong to the same community. The others belong to different communities.

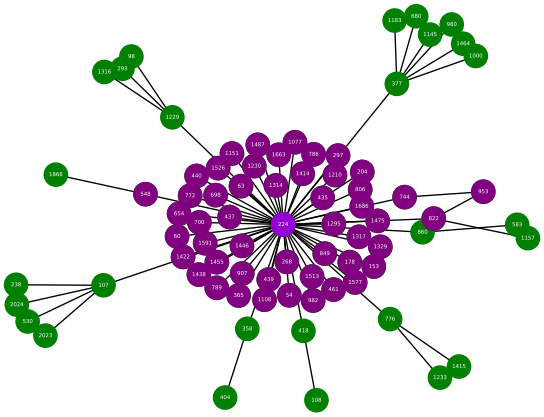


**Figure B.2:** A 2-hop ego network of node 224 of the Yeast Protein network. This node has the maximum degree in this network (56). Nodes with the color purple belong to the same community as 224 . The others belong to different communities.

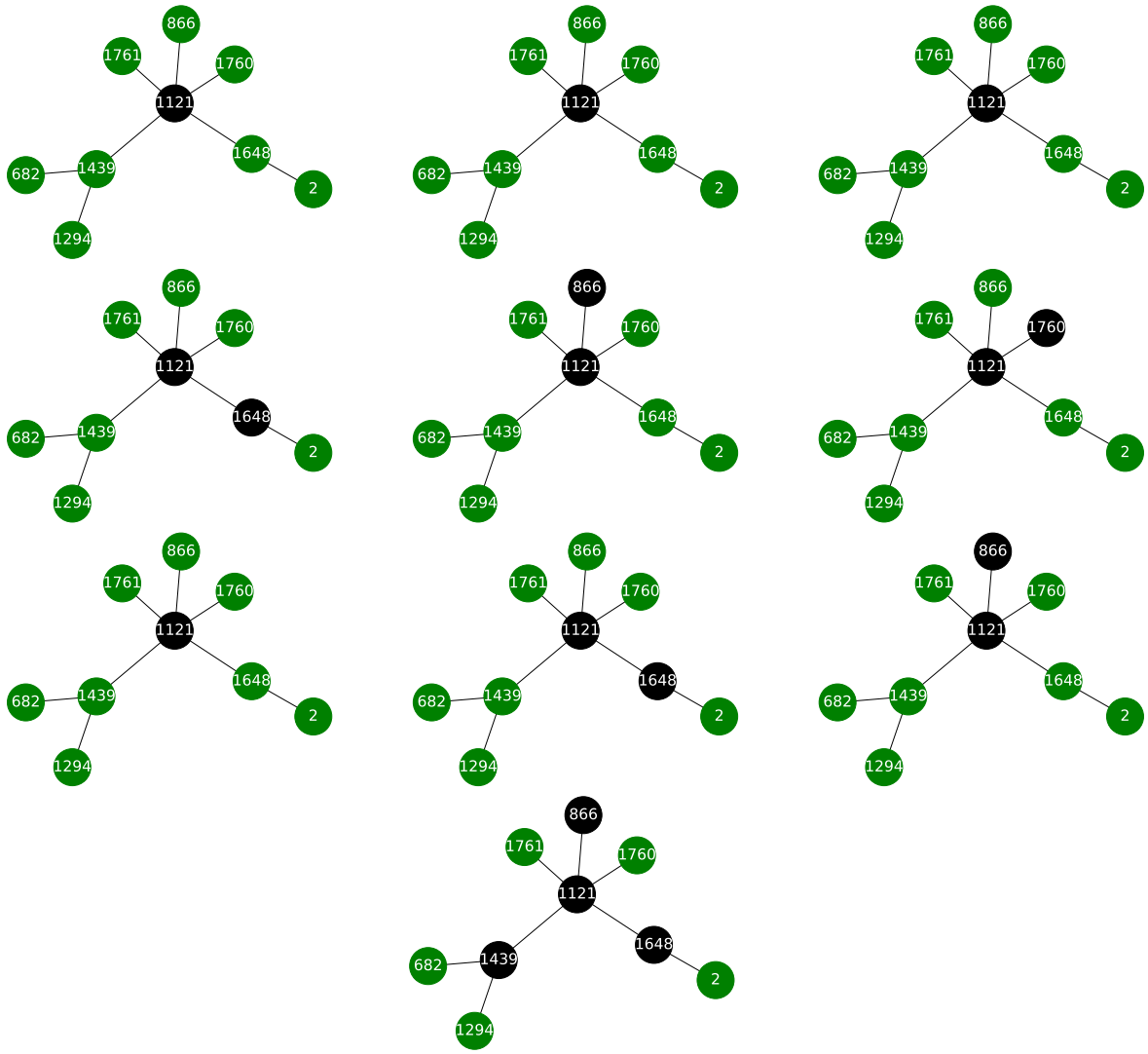



**Figure B.3:** 10 SIR simulations with 1121 as the seed node, at the end of the infection spreading process at $\lambda = \lambda_c$. Nodes in black are nodes that have recovered. Nodes in green were not infected.

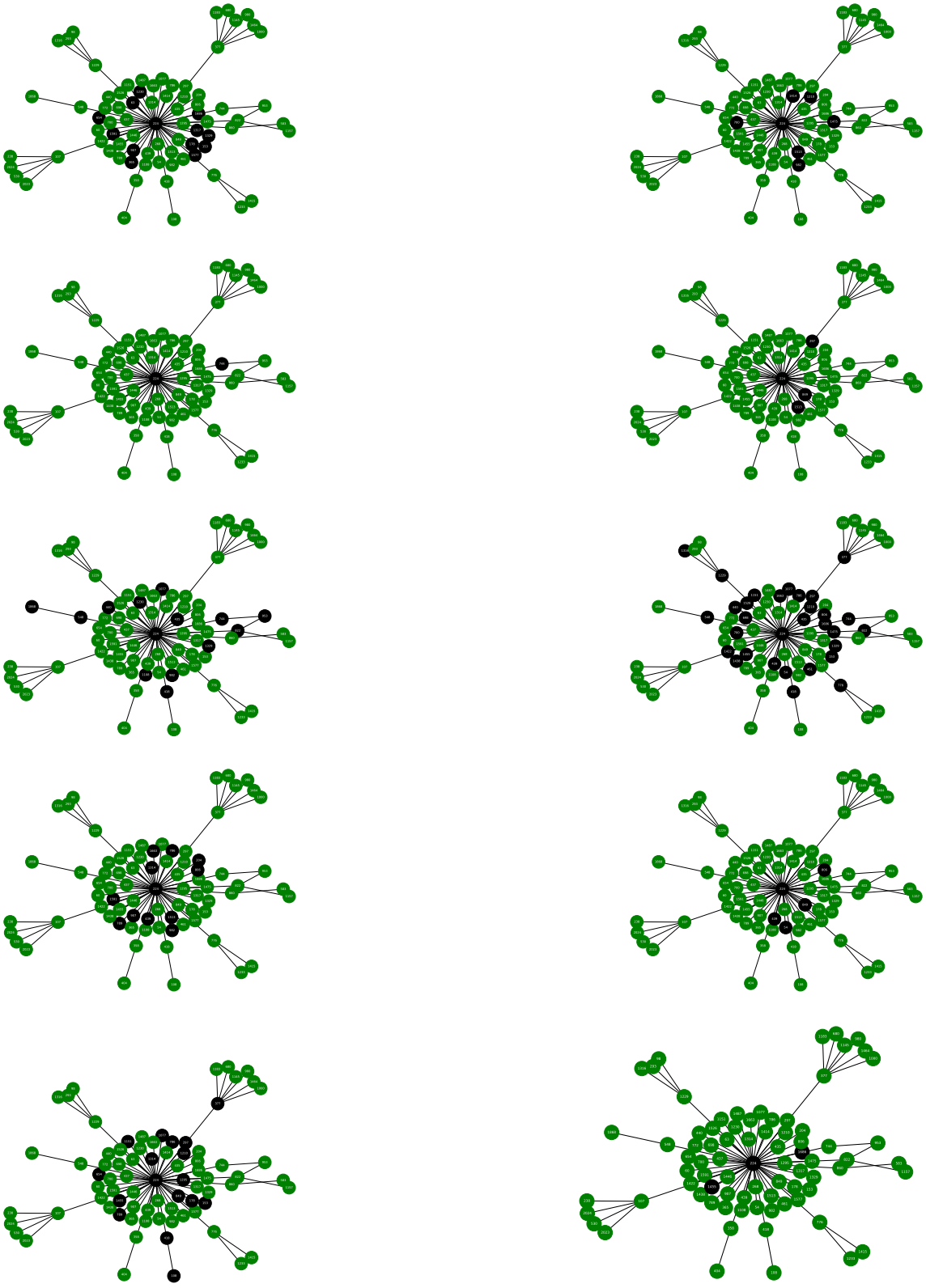
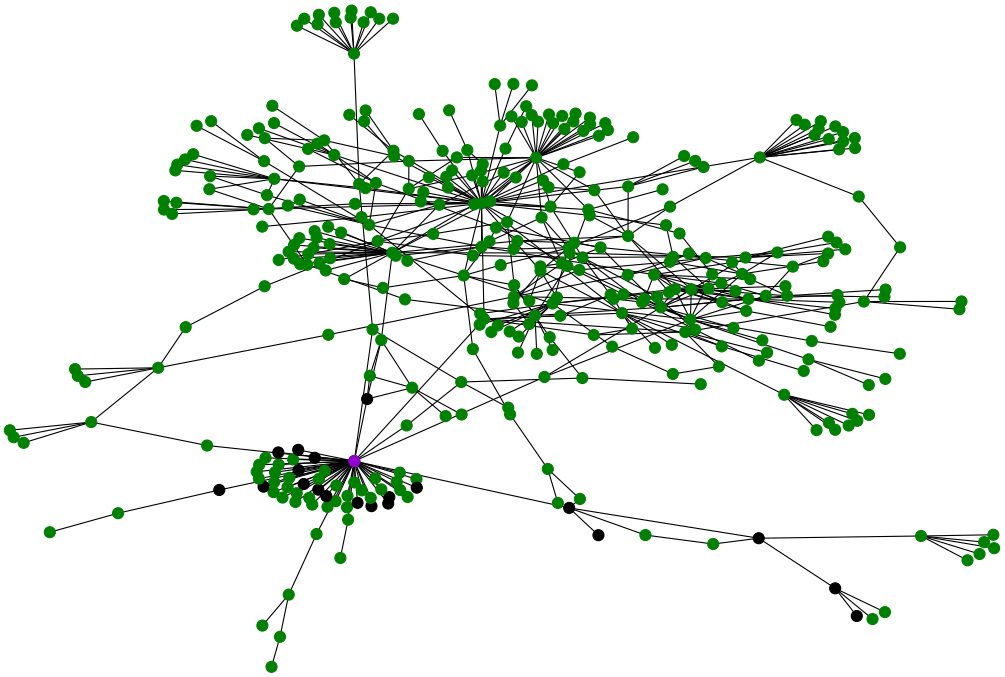
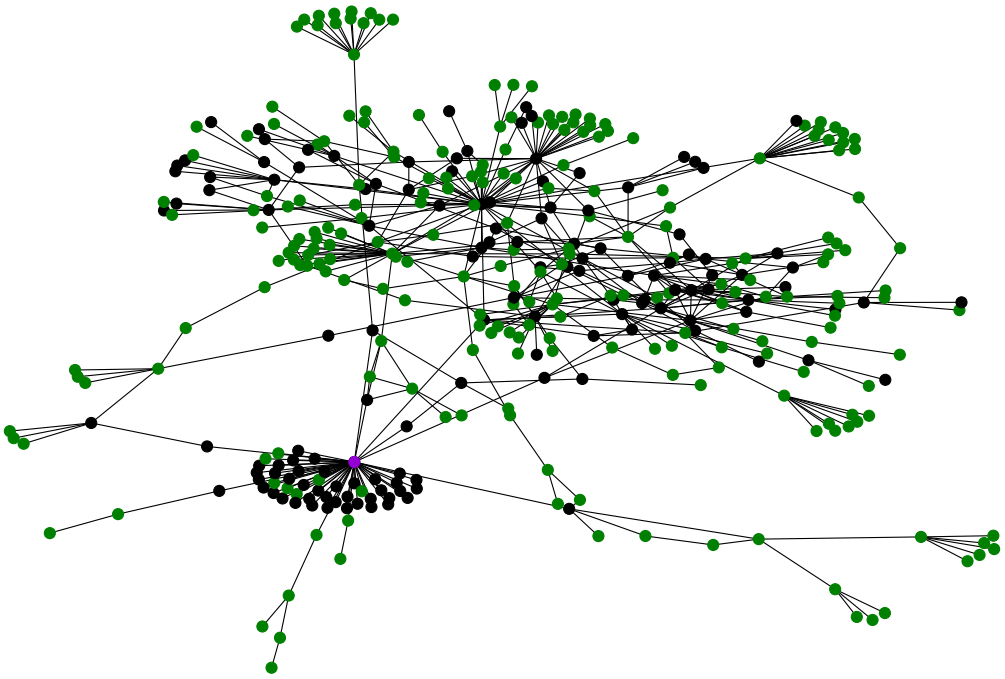


**Figure B.4:** 10 SIR simulations with 224 as the seed node, at the end of the infection spreading process. $\lambda = \lambda_c$. Nodes in black are nodes that have recovered. Nodes in green were not infected. Seven out of ten times, the infection spreads to only one-hop neighbors.



**(a)** Simulation at $2\lambda_c$



**(b)** Simulation at $3\lambda_c$

**Figure B.5:** SIR simulations with 224 as the seed node (in purple), at the end of the infection spreading process. The 4-hop neighborhood of 224 is considered. 10 simulations are run for each effective transmission rates - $2\lambda_c$ and $3\lambda_c$. The run with the highest outbreak size among the 10 runs is shown in this figure. At $2\lambda_c$, the infection can spread beyond 2-hop neighborhood, but the number of infections in such neighborhoods is relatively small. At $3\lambda_c$, infection can spread beyond 2-hop neighborhood of the seed node with large number of recovered nodes.