# The Impact of Interaction and Algorithm Choice on Identified Communities

Rana Maher*, David Malone†
Hamilton Institute
Maynooth University, Kildare, Ireland
*rana.maher.2016@mumail.ie, †david.malone@nuim.ie

Marie Wallace
Emerging Technology Center
IBM Analytics Group, Dublin, Ireland
marie.wallace@ie.ibm.com

***Abstract*—In social networks, nodes are organized into densely linked communities where edges appear among the nodes with high concentration. Identifying communities has proven to be a challenging task due to various community definitions/algorithms and also due to the lack of "ground truth" for reference and evaluation. These communities not only differ due to various definitions but also can be affected by the type of interactions modeled in the network, which lead to different social groups. We are interested in exploring and studying the concept of *partial network views*, which is based on multiple types of interactions. An *Enron* email network is used to conduct our experiments. In this paper, we explore the mutual impact of selecting different views extracted from the same network and their interplay with various community detection algorithms to measure the change and the level of realism of the structure for non-overlapping communities. To better understand this, we assess the agreement of partitions by evaluating the partitioning quality (performance) and finding the similarity between algorithms. The results demonstrate that the topological properties of communities and the performance of algorithms are equivalent to each other. Both of them are affected by the type of interaction specified in each view. Some network views appeared to have more interesting communities than other views, thus, might help to approach a relatively informative and logic "ground truth" for communities.**

***Index Terms*—Community detection; Community structure; Similarity measures; Partitions comparison**

## I. Introduction

In the real-world, similarities or connections between entities can be determined by various relationships/interactions. Relationship can be friendship between actors (e.g., family, business, school) or how they communicate (e.g., email, mobile, text). These kind of relations in social networks are represented in what is called a "social graphs", where edges here represent different types of relationships between actors. For example, in the Enron email network, different relationships/interactions can be who cc'd, sends to or bcc'd who in the network, which represent the type of the exchanged emails in the network. Deriving different communication networks or graphs from the same network, where each one corresponds to a specific type of interaction, can affect the derived communities. We will denote to these interaction-based graphs in our study as *partial network views* where each view models a different and specific interaction within the network.

Community detection in social graphs is one of the considerable interests and challenges that have acquired great attention (e.g. [1]). In fact, it has been targeted by many studies and considered as an important part in the area of *Network Analysis*. The reason for this is that communities reveal necessary understanding for analyzing the behavior of people with each other within the network. In networks/graphs, a community can be defined as a group of nodes that have a better connection within a group and sparsely connected with other groups in the network or in other words, nodes that have better internal connections than the external ones [1]. Therefore, algorithms exist to detect communities which are denser in connection, smaller in structure and have strong connections within its vertices. Generally, most communities which have dense links are most likely to have common properties, that is why the concept of similarity is linked to the concept of community. Hence, similarity measures have a great influence on detecting communities [2].

There are various definitions for community detection algorithms that have been presented and studied in the research community. Each algorithm has its own approach in defining communities. Some of the different definitions are: random walks, spectral analysis, label propagation, centrality, modularity and many others [3]. These algorithms can be compared from different perspectives, either from the process that lead to finding the community structure or from studying the community structure itself. Choosing the best algorithm that can suit specific problem is not an easy task [4].

In this work we focus on studying and exploring three questions: How similar are the results of different community detection algorithms? How does the partial network views concept affect the derived communities? Which type of network view lead to more informative communities? Generally, we further study the change in communities under the change of two factors: first, the partial network views and second, the community detection algorithms. The task of comparing sets of communities has various applications, one of the obvious ones is to compare the results with the "ground truth". Another approach is to compare the results of different community algorithms to each other. It is worth emphasizing that we can not compare communities to "ground truth" as Enron data lacks the "ground truth" for communities. However, our main contribution is to study these questions and explore how social

community structure and the quality of partitioning change over various network views and algorithms. We will do this by quantifying their structural properties and applying different quality measurements to elucidate the issue of the community structure differences over multiple views.

The paper is organized as follows: an overview of community detection approaches are presented in Section II. Some measures/criteria that have been proposed in the literature to reflect the properties of community structure are presented in Section III. Evaluation measures for comparing the quality of partitions and their similarity are discussed in Section IV. An overview of the enron dataset and clarification of our approach in refining and modeling the data is proposed in Section V. Then, the results and the discussion are provided in Section VII and Section VIII respectively. Section IX concludes the paper and highlights the future work.

## II. Community Detection Approaches

An informal definition of the community as a group with densely connected edges has been common. However, there are numerous definitions or approaches for algorithms that implement different strategies for finding the community structure. Approaches can be based on optimizing modularity, like FastGreedy [5], Louvain [6] and Spinglass [7], or algorithms like Leading Eigenvector [8] and Commfind [9], which are spectral algorithms. Others algorithms that are based on random walks like MarcovCluster [10] and Walktrap [11], or information theoretic algorithms, like Infomod [12] or Infomap [13]. In our study, we select some algorithms from four main categories that differ in the approach of identifying communities.

*Walktrap* (WT) is a random walk based algorithm where hierarchical agglomerative clustering is the applied approach proposed by Pons and Latapy [11]. Random walks mean that at each step the algorithm moves from one node to another through a random choice. Generally, the idea is to use the distance measure from one node to another to identify communities. For example, if two nodes $j$ and $k$ are in same community, then the probability to a third random node $i$ to be in the same community should not be that different from both $j$ and $k$. This algorithm uses a node similarity approach where each community is detected as a group of nodes similar to each other and dissimilar from the rest of the nodes in the network.

*Infomap* (IM) is an example of compression-based approach. It is based on information theoretic principles. The community structure here is derived based on Huffman coding [13]. It tries to minimize or compress the information quantity over the network. This approach does not use the separation and cohesion concepts like other community detection definitions.

*Label Propagation* (LP) uses the neighborhood concept and depends on assigning each node to one unique label from $k$ labels. Then an iterative process takes place where each node is assigned the label that is mostly common in its neighborhood. The process stops when each node has the label that is the most frequent in its neighborhood. Communities are then constructed by targeting groups of nodes having the same label [14]. These type of algorithms requires no prior information or parameter settings. The only guide here is the network structure.

*Edge Betweenness* (EB) is proposed by Girvan and Newman [15]. It is a hierarchical process where the edges are removed from the network in a decreasing order according to their edge betweenness scores. It measures the centrality of a specific edge by finding the percentage of shortest paths passing through this edge in the network and this highlights the importance of the edge. The algorithm yields a good results but is not commonly used for large-scale graphs due to its high computational complexity. This type of algorithms is based on measures of centrality approaches. They deal with the network as one entity where the network splits into multiple components by repeatedly removing the central edges.

## III. Topological characteristics of communities

In this section, we study some popular topological properties of communities to discuss in later sections how far these properties can be affected by different partitioning on different network interactions.

### A. Size

The size of the community is the overall number of communities including communities with single vertex derived by a community detection algorithm.

### B. Size distribution

The distribution of the community size is one of the important features of the community structure. They are often unevenly distributed and sometimes obey a power law [16] with exponent ranging between 1 to 2 [17]. The minimum size distribution of a community in real networks is 2 while the maximum size can vary in a wide range depending on the used model [18].

### C. Singleton

A singleton community is the community that contains only a single vertex.

### D. Transitivity

The transitivity depends on how the direct neighbors of a certain node are connected. It is the actual number of links (edges) between neighbors, divided by all the possible links if they are all connected. The internal transitivity $T$ of a community $C$ is defined as:

$$T = \frac{1}{n_C} \sum_{i \epsilon C} \frac{2 * l(i)}{k_{int}(i)[k_{int}(i) - 1]} \quad (1)$$

where $n_C$ is the number of nodes in community $C$, $l(i)$ is the number of actual links between neighbors of the node $i$ and $K_{int}(i)$ is the internal degree of the node $i$.

### E. Edge Density

The edge density $\mathcal{P}$ of a community $C$ of an unweighted graph is the ratio between the actual realized links in the community $E_c$ to the maximum number of possible links it can contain if all the nodes are well connected to each other $N_C(N_c-1)/2$ where $N_c$ is the number of nodes in the community. Communities are supposed to be higher in density than the whole network. The density function is defined as:

$$\mathcal{P} = \frac{2|E_c|}{|N_c|(|N_c|-1)} \tag{2}$$

## IV. Evaluation of community detection algorithms

Evaluating and comparing partitions is a classical problem in the area of community detection. But, since there is no perfect single quality measure for comparing communities from different algorithms [19]. Therefore, in this section we will review some of the most commonly used measures that have been presented in the literature. Some of these measures are based on global metric like *Modularity* which compares the results relative to random graphs or based on counting pairs like *Random Index* or *Adjusted Random Index* and others are based on the use of mutual information like *Normalized Mutual Information*. The general strategy for comparing partitions as shown in Figure 1 where $P_1 = C_1^{'}, C_2^{'}, ..C_n^{'}$ and $P_2 = C_1^{''}, C_2^{''}, ..C_m^{''}$ are examples of two sets of communities from different partitions.

### A. Modularity

Modularity is one of the most commonly used method to evaluate the quality of partitioning a network into communities [20]. It is used to measure the quality of division within a network into communities. It is a common measure used to determine and compare the performance of community detection algorithms. The value of modularity ranges between $-1$ and $1$, a higher value means better partitioning. This function is defined in [21] as:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{K_i K_j}{2m}\right)\delta(c_i, c_j) \tag{3}$$

where $A_{ij}$ is the adjacency matrix of the network between node $i$ and $j$, $K_i$ is the degree of node $i$ and $c_i$ is the community index of node $i$ and $\delta(u,v) = 1$ if $u = v$ and $\delta(u,v) = 0$ if otherwise.

### B. Random Index (RI)

The Random Index measures the agreement for a given pair of nodes to be in same community for the estimated and reference partition. It counts the pairs of nodes that are classified correctly. The $RI$ ranges from $0$ (when pair misclassified) and $1$ (correctly classified) under different partitions [22]. The $RI$ is defined as follows:

$$R(P_1, P_2) = \frac{2(n_{11} + n_{00})}{n(n-1)} \tag{4}$$

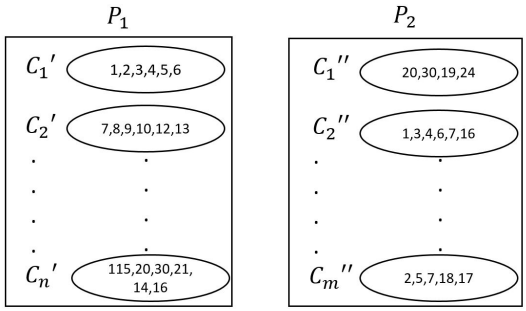



Figure 1: Non-overlapping communities derived from different community detection algorithms on the same data.

The value of $R$ depends on both the number of nodes and the number of partitions. $P_1$ and $P_2$ are two different partitions, $n$ is the number of nodes, $n_{11}$ is the number of nodes that grouped in the same community for partitions $P_1$ and $P_2$ while $n_{00}$ is the number of nodes that grouped in different community for both partitions.

### C. Adjusted Random Index (ARI)

This is an adjusted version of RI that is proposed by Hubert and Arabie [23]. The $ARI$ is the normalized difference of $RI$ and the value expected under a null hypothesis (the number of clusters be the same in the two clustering). It compares how much two partitions have common information between each other using the same dataset. The measure takes the value $1$ when the resulting partition perfectly matches the reference one and takes value $0$ when the algorithm fails completely to detect the appropriate community structure.

### D. Normalized Mutual Information (NMI)

$NMI$ is one of the classical measures that ranges from $0$ to $1$ when perfectly corresponds to the reference partition. It compares how much common information between two different partitions. It is used in the research community [24] to measure the performance of community detection algorithms.

## V. Enron Data

### A. Enron Background

Enron is a dataset that has large number of emails for individuals of Enron's staff. It is a fertile source for a real corporation within the period 1998–2002. The data was made publicly available by the *Federal Energy Regulatory Commission*. William Cohen made the dataset available public on-line for researchers. Enron data has different types of emails either personal or official; the dataset that is on-line contain around $151$ employees and all emails between them in $3500$ folders. It is organized as folder for each employee and inside each folder sub-folders for the emails sent, deleted and junk. The staff are mostly from the management level starting from CEO to Vice President. The version which we will conduct our research on is based on the MySql database that was formulated by Shetty and Adibi [25]. The database consists of four tables, each one represents different entity like: employees, messages,

recipients and references. They cleaned the emails by deleting any unneeded ones, duplicates or even blank ones and fixing aliasing problems.

### B. Enron extraction and refinement

The database represents two types of information; the first is the communication type between the employees and what we mean by communication type is the *TO*, *CC* or *BCC* fields, the second type is the content of the messages in the emails. Our work will focus on the first type of information. We focus on extracting the communications between the employees. We deleted all personal emails and any emails outside the organization. We dropped any emails ending with enron.com but not within the 151 employees.

## VI. Different Network Views extraction

Our focus is to derive three different communications network views as we illustrated before. Nodes will represent Enron employees and edges between them correspond to a specific type of communication which will differentiate between the views. We derived all the communications between the Enron employees and then filtered them according to communication type. Our focus is on extracting the *TO* and the *CC* fields from the Enron dataset and model each one as a graph where each corresponds to a partial view in addition to a general view that only indicates that there is an existing edge whatever the type of communication.

Hence, the first view will represent an *undirected graph*, we refer to this network as *Netview*, a single edge will exist here between any two emails if there is any type of exchange in emails between them. The second view is based on the $TO$ field in the emails, we represent this as a *directed graph* that reflects the direction of communication of the emails, we denote to this graph as *TO Netview*. For the third view, edges correspond to the *CC* field in emails, we represent a *directed graph* and refer to it as *CC Netview*. We note that we were interested in studying the *BCC* communications but these were not available in the dataset.

As shown in Table I, we explored some generic properties for each view like: *nodes*, *edges*, *cliques* (nodes that are tightly connected to each other) and *clustering coefficient* (the degree of nodes that tend to cluster with each other). We found that an employee named " Paul Barbo" is missing from the three network views, also, for the *CC Netview* around six employees do not appear to have any communication (Mary Fischer, Steven Merris, Joe Stephenovitch, Joe Quenet, Andrew Lewis, Paul Barbo). The three networks views have almost the same clustering coefficient ($\approx 0.3$). The number of cliques is bigger in the undirected network (*Netview*) and remains the same for both directed networks (*TO Netview*, *CC Netview*) although they are quite different in the number of edges.

## VII. Results

After extracting different partial network views from the enron dataset, as discussed in the previous section, the four different community detection algorithms which are presented in Section II are used to study the level of realism of the resulting communities. We studied how each algorithm performs differently on each view from two perspectives: (a) Topological characteristics of the resulting communities as described in depth in Section VII-A. (b) Partitioning quality of each of the algorithm's output as illustrated in Section VII-B. Our main focus is to compare the results of the algorithms with each other and not to the unavailable "ground truth" as highlighted in Section I. We stress on the fact that the community detection algorithms in most cases lead to different results whether it is from the partitioning point of view or the topological characteristics. Hence, we evaluated the resulting communities from both prospectives to complement each other.

| Network Views | Netview | TO Netview | CC Netview |
|---|---|---|---|
| Nodes | 150 | 150 | 145 |
| Edges | 1511 | 2007 | 799 |
| Cliques | 12 | 8 | 8 |
| Clustering Coefficient | 0.388 | 0.37 | 0.33 |

Table I: General properties for the three communication network views from Enron network.

| Network Views | Netview | TO Netview | CC Netview |
|---|---|---|---|
| IM | 7 | 9 | 11 |
| WT | 4 | 9 | 15 |
| LP | 4 | 3 | 9 |
| EB | 21 | 9 | 55 |

Table II: Different community sizes generated from the algorithms for three different communication network views including the singleton communities.

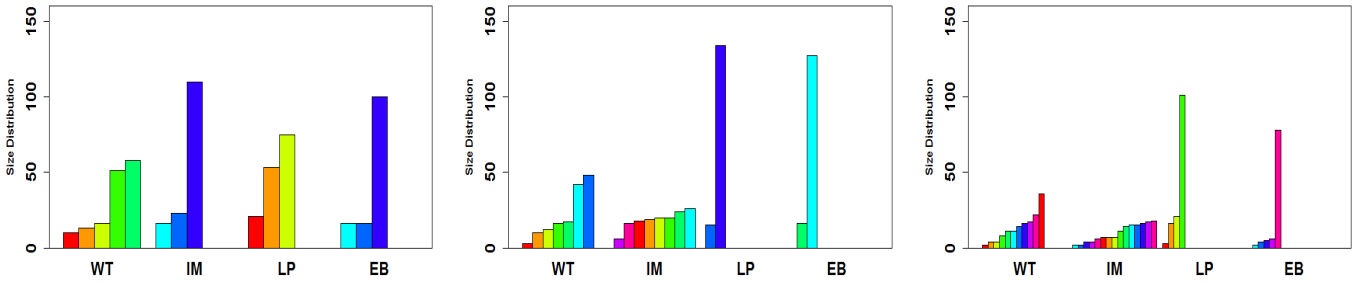
### A. Topological Characteristics

This section shows how the four community detection algorithms with the three different network views can impact the topological properties of the resulting communities.

*1) Size, Size Distribution & Singleton:* The different community *size* resulted from the various algorithms are highlighted in Table II. Figure 2 illustrates the *size distribution* for each of the three different network views with an exception of the singleton communities; where the *x-axis* represents the algorithm and the *y-axis* shows the *size distribution*. While, the *singleton* communities appear in Table III. It is worth emphasizing the following:

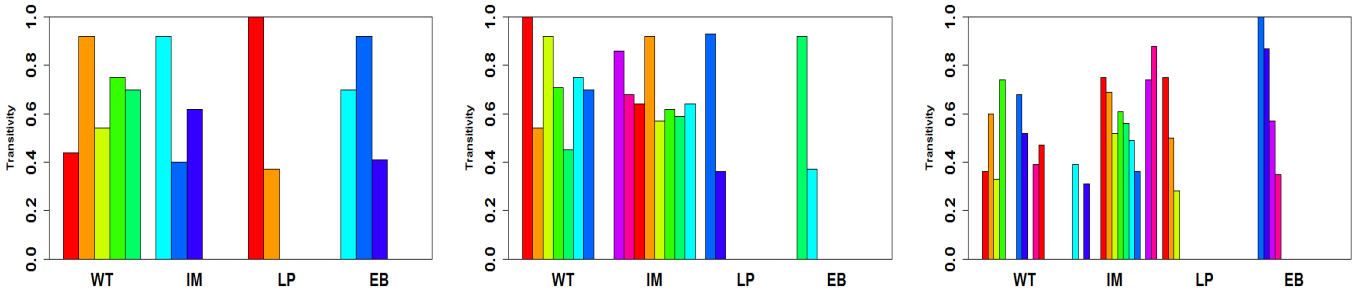
| Network Views | Netview | TO Netview | CC Netview |
|---|---|---|---|
| WT | 2 | 2 | 0 |
| IM | 1 | 1 | 0 |
| LP | 1 | 1 | 4 |
| EB | 18 | 7 | 48 |

Table III: The number of singleton communities generated from the algorithms for three different communication network views.



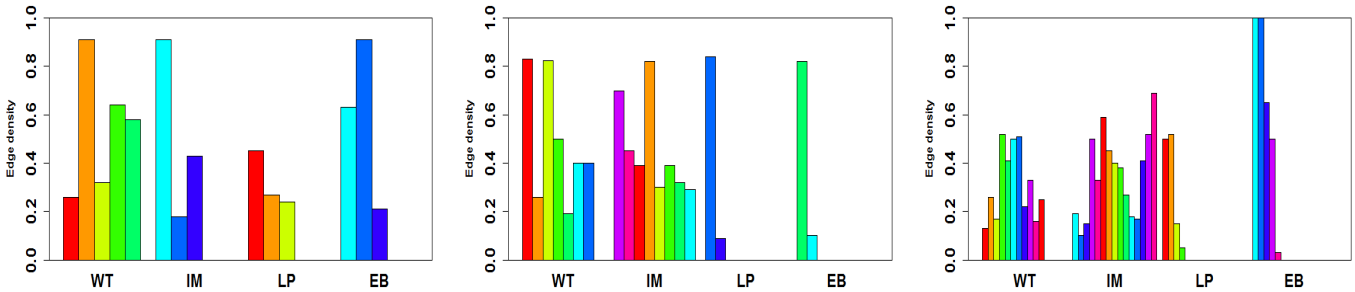

(a) Netview (b) TO Netview (c) CC Netview

Figure 2: Size distribution plotted in bars to represent the measures for each community (*y-axis*) derived from each of the four algorithms (*x-axis*) for *Netview*, *TO Netview* and *CC Netview*. The plotting here considers communities resulting from each algorithm except the singleton communities for 150 employees in Enron network.




(a) Netview (b) TO Netview (c) CC Netview

Figure 3: Transitivity plotted in bars to represent the measures for each community (*y-axis*) derived from each of the four algorithms (*x-axis*) for *Netview*, *TO Netview* and *CC Netview*. The plotting here considers communities resulting from each algorithm except the singleton communities for 150 employees in Enron network.




(a) Netview (b) TO Netview (c) CC Netview

Figure 4: Edge Density plotted in bars to represent the measures for each community (*y-axis*) derived from each of the four algorithms (*x-axis*) for *Netview*, *TO Netview* and *CC Netview*. The plotting here considers communities resulting from each algorithm except the singleton communities for 150 employees in Enron network.

- EB shows the poorest performance across all views specially for the *CC Netview*. This can be observed from the number of communities having more than one vertex (3, 2, and 5 communities in the *Netview*, *TO Netview*, and *CC Netview* respectively) with respect to the total number of the resulted communities, it gives a large number of singleton communities for the three views (See Table II and Table III).
- WT and IM perform reasonably well for both the *TO Netview* resulting in 7 and 8 communities more than one vertex respectively. Similarly for the *CC Netview* which resulted in 10 and 15 communities respectively. Both perform perfectly with the *CC Netview* deriving 0 singletons.
- For LP, the communities resulted for the *CC Netview* is more stable compared to the *Netview* and *TO Netview*. This is illustrated from the ratio between the number of communities that contain more than one vertex (4 communities) with respect to the total number of communities (9 communities) corresponding to $4/9 \approx 0.4$ compared to the 0.75 and 0.67 ratio resulted for the *Netview* and *TO Netview* respectively. For the singletons, interestingly, it extracts a less number of singletons than WT for *Netview* and *TO Netview*.

We could claim intuitively that the lower the ratio between the number of communities that contain more than one vertex with respect to the total number of communities, the more uniform distributed communities and hence, a higher chance to derive an interesting insights from such communities.

*2) Transitivity:* Figure 3 shows the transitivity for the four algorithms, which show different trends. We observed that:

- For LP and EB, we observe that the larger community size, the higher transitivity and vice versa. This inverse relation is found to be consistent in all network views.
- It is worth noticing that despite the fact that EB resulted in a poor partitioning (based on the size distribution property), some of the resulting communities score high transitivity ($> 0.8$).
- For IM and WT, the *TO Netview* is observed to have high transitivity values compared to the *CC Netview* ($> 0.5$) for most communities. Communities which score transitivity equals to 1 is mainly comprised of 2 or 3 nodes which follows the intuitive thinking expected from very small communities.

*3) Edge density:* As shown in Figure 4, some of the observed interesting points are:

- For LP and EB, a defined relation between edge density and community size; edge density decreases with the increase of community size and vice versa resulting in large sparse communities and dense small size communities.
- For IM and WT, there is slightly big difference in edge density values for communities in the *Netview*. This difference decreases in the *TO Netview* and further decreases in the *CC Netview*; communities tend to get close in their edge density values which reflects that they are more homogeneous (all of them range between 0.2 to 0.6).

| Network Views | Netview | TO Netview | CC Netview |
|---|---|---|---|
| WT | 0.35 | 0.4 | 0.55 |
| IM | 0.23 | 0.4 | 0.54 |
| LP | 0.11 | 0.15 | 0.43 |
| EB | 0.2 | 0.17 | 0.145 |

Table IV: The quality of partitioning expressed in terms of modularity function for three different communication network views.

### B. Evaluation of Partitions

Comparing two sets of partitions is not an easy task and can be achieved by different ways. We can either compare the resulting communities to the ground truth or to the results of other community detection algorithms. Our comparison is based on the second type. We compare the partitioning extracted from algorithms for each view to check the agreement of the existing a set of nodes in one community (or in different one). Hence, the agreement corresponds to the proportion of a set of nodes for which two communities agree across each network view. This evaluation is performed using the quality measures discussed in Section IV.

*1) Modularity:* is used to measure the quality of partitioning in general, as shown in Table IV to find out which is the best partitioning in each case. We found that:

- IM and WT have the highest modularity when compared to LP and EB. Both IM and WT score very close modularity index among all views.
- The quality of partitioning for all the algorithms is higher in the *CC Netview* than the *TO Netview*.
- The quality of partitioning of the *TO Netview* is higher than the *Netview* except for the EB.

*2) Similarity measure:* is used to compare the resulted partitions to highlight their similarity level. When looking at RI, ARI and NMI in Table V, Table VI and Table VII respectively. We observed that the three similarity measures agree on the order of the similarity values for both the views and the algorithms. However, values of ARI are higher than NMI values which are higher than RI. Our discussion is valid for any of the three measures.

- EB & WT/IM/LP, although EB nearly scores the least modularity for all the network views across the other three algorithms but it shows a high similarity score when compared to them specially for the *Netview*. Generally, these high values do not match with the high divergence in the community sizes distribution. The only view that matches with that divergence is the *CC Netview* and gives low similarity scores as well, where the performances observed to be very low (nearly close to zero) for ARI measure with other algorithms.
- IM & WT score the highest similarity measure when compared to each other for the *CC Netview* followed by

| View | Netview | | | | TO Netview | | | | CC Netview | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RI | **IM** | **LP** | **WT** | **EB** | **IM** | **LP** | **WT** | **EB** | **IM** | **LP** | **WT** | **EB** |
| **IM** | 1.000 | 0.709 | 0.712 | 0.876 | 1.000 | 0.319 | 0.865 | 0.380 | 1.000 | 0.837 | 0.915 | 0.704 |
| **LP** | 0.709 | 1.000 | 0.423 | 0.603 | 0.319 | 1.000 | 0.401 | 0.917 | 0.837 | 1.000 | 0.857 | 0.636 |
| **WT** | 0.712 | 0.423 | 1.000 | 0.720 | 0.865 | 0.401 | 1.000 | 0.459 | 0.915 | 0.857 | 1.000 | 0.681 |
| **EB** | 0.876 | 0.603 | 0.720 | 1.000 | 0.380 | 0.917 | 0.459 | 1.000 | 0.704 | 0.636 | 0.681 | 1.000 |

Table V: *Random Index (RI)* values represented to find the best two matching algorithm or *Netview*, *TO Netview* and *CC Netview*.

| View | Netview | | | | TO Netview | | | | CC Netview | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ARI | **IM** | **LP** | **WT** | **EB** | **IM** | **LP** | **WT** | **EB** | **IM** | **LP** | **WT** | **EB** |
| **IM** | 1.000 | 0.355 | 0.457 | 0.756 | 1.000 | 0.064 | 0.530 | 0.072 | 1.000 | 0.358 | 0.550 | 0.091 |
| **LP** | 0.355 | 1.000 | 0.114 | 0.245 | 0.064 | 1.000 | 0.116 | 0.770 | 0.358 | 1.000 | 0.492 | 0.033 |
| **WT** | 0.457 | 0.114 | 1.000 | 0.422 | 0.530 | 0.116 | 1.000 | 0.143 | 0.550 | 0.492 | 1.000 | 0.077 |
| **EB** | 0.756 | 0.245 | 0.422 | 1.000 | 0.072 | 0.770 | 0.143 | 1.000 | 0.091 | 0.033 | 0.077 | 1.000 |

Table VI: *Adjusted Random Index (ARI)* values represented to find the best two matching algorithm or *Netview*, *TO Netview* and *CC Netview*.

| View | Netview | | | | TO Netview | | | | CC Netview | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NMI | **IM** | **LP** | **WT** | **EB** | **IM** | **LP** | **WT** | **EB** | **IM** | **LP** | **WT** | **EB** |
| **IM** | 1.000 | 0.432 | 0.689 | 0.701 | 1.000 | 0.281 | 0.739 | 0.339 | 1.000 | 0.642 | 0.796 | 0.529 |
| **LP** | 0.432 | 1.000 | 0.271 | 0.285 | 0.281 | 1.000 | 0.323 | 0.694 | 0.642 | 1.000 | 0.667 | 0.411 |
| **WT** | 0.689 | 0.271 | 1.000 | 0.629 | 0.739 | 0.323 | 1.000 | 0.404 | 0.796 | 0.667 | 1.000 | 0.485 |
| **EB** | 0.701 | 0.285 | 0.629 | 1.000 | 0.339 | 0.694 | 0.404 | 1.000 | 0.529 | 0.411 | 0.485 | 1.000 |

Table VII: *Normalized Mutual Information (NMI)* values represented to find the best two matching algorithm or *Netview*, *TO Netview* and *CC Netview*.

the *TO Netview* and that means that their similarity value affected significantly by the type of the relationship in the graph. For example, their values reached $0.9$ for the RI measure for the *CC Netview* and this agreed with their corresponding modularity values which is more than $0.5$.

- LP & WT obtained the highest values for the *CC Netview* and then the *TO Netview.*
- Most of the algorithms showed better value for the *TO Netview* more than the *Netview* except LP & IM, both of them give higher similarity over the *Netview* than the *TO Netview*.

## VIII. Discussion

*From the topological properties point of view*, *size distribution* got affected by both algorithms and the type of the network view. It is clear that changing the views do not impact the rate of *singletons* but the algorithm itself is the main impact on their rate. The *transitivity* and the *edge density* are not always dependent on the size of communities. The partitioning of *CC Netview* with IM and WT worth having an attention as it may give insights that might be close to real-world. Directed networks can be more fertile for partitioning and insightful than undirected ones even if they are denser (e.g. *TO Netview* better than *Netview*). Some directed networks can be more interesting than other directed ones according to the modeled relation type. We can infer this from the *CC Netview* where the derived communities seem to be more appropriate than the *TO Netview* as in fact they correspond to a more obvious relationship between any two employees. Interestingly, we found some communities in the views act as a sub-network and tend to split into communities in other views. For instance, in WT, some communities in the *Netview* get split in the *TO Netview* and communities in the latter one tends to split in the *CC Netview*. Hence, it is clear that deciding whether the derived communities are well formed or it could be split further into communities, is not an easy task.

*From the quality of partitioning point of view*, It is clear that among the three network views that WT & IM have the highest scores. The use of different similarity measures did not matter much for our data, this could save time in evaluations. It is observed that the highest scores always achieved for the *CC Netview* and *TO Netview* across most of the measures. This can indicate that: the more relationship is specified in the graph, the higher the quality of the partitions than modeling a graph without specific relationship. It is also clear that the quality results coincide with the topological results, both of them agree that IM is almost as good as WT especially for the *CC Netview* which has been shown to have better topology structure and higher quality of partitioning as well. As the *CC Netview* scores higher results than the *TO Netview* and the latter scores higher results than the *Netview*, therefore, we expect that we can derive more interesting communities when modeling the *BCC* communications.

Generally, we found that the quality of partitioning and the topological structure are equivalent to great extent, however, it was hinted in the literature that they are not equivalent to each other [3]. For instance, the equivalence is clearly defined for Infomap and Walktrap. We found that the graph based on specific relationship between its nodes in the Enron network play an observable role, it is obvious that modeling a graph based on one type relationship can have an observable impact on optimizing the structure and the quality of communities. Generally, the view can have a big impact, as can the algorithm.

## IX. Conclusion and Future work

The aim of this paper is to study the entire derived communities based on different relationships/interactions representation for the Enron email network using four different community detection algorithms such as: Walktrap, Infomap, Label propagation and Edge Betweenness. The goal from this study is to discover how the topological properties and the quality of partitioning of communities got affected by two factors: various network interactions (partial network views) and multiple community detection approaches. We aim that this study tackle the idea of selecting an interesting interaction and a reliable algorithm for detecting communities which has been a challenging task with the lack of "ground truth" for many networks nowadays.

We studied some topological properties of the estimated communities resulted from each algorithm applied on three network views. We then asses the results through evaluating the quality in terms of comparing partitions using modularity measure and other similarity measures. The considered measures mostly agree with each other on the quality of partitioning with tiny differences and equivalent some how to the topological properties.

This work can be extended to a more comprehensive analysis by expanding the properties and using a wide range of algorithms with some additional community structure measures. It would be also useful if we consider other email networks and observe if they behave similarly or not. Since each of the community detection algorithms based on optimizing only one criteria, future work can be to diffuse different community detection approaches with multiple criteria to a single algorithm in order to obtain "Robust Communities".

## Acknowledgment

This publication emanated from research supported in part by a grant from Science Foundation Ireland (SFI) and co-funded under the European Regional Development Fund under Grant Number 13/RC/2077.

## References

[1] J. Leskovec, K. J. Lang, and M. Mahoney, "Empirical comparison of algorithms for network community detection," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 631–640.

[2] W. Fan and K. Yeung, "Similarity between community structures of different online social networks and its impact on underlying community detection," *Communications in Nonlinear Science and Numerical Simulation*, vol. 20, no. 3, pp. 1015–1025, 2015.

[3] G. K. Orman, V. Labatut, and H. Cherifi, "Comparative evaluation of community detection algorithms: a topological approach," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2012, no. 08, p. P08001, 2012.

[4] L. Peel, "Estimating network parameters for selecting community detection algorithms," in *Information Fusion (FUSION), 2010 13th Conference on*. IEEE, 2010, pp. 1–8.

[5] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical review E*, vol. 69, no. 6, p. 066133, 2004.

[6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[7] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," *Physical Review E*, vol. 74, no. 1, p. 016110, 2006.

[8] M. E. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical review E*, vol. 74, no. 3, p. 036104, 2006.

[9] L. Donetti and M. A. Muñoz, "Improved spectral algorithm for the detection of network communities," *arXiv preprint physics/0504059*, 2005.

[10] S. Van Dongen, "Graph clustering via a discrete uncoupling process," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 1, pp. 121–141, 2008.

[11] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *International Symposium on Computer and Information Sciences*. Springer, 2005, pp. 284–293.

[12] M. Rosvall and C. T. Bergstrom, "An information-theoretic framework for resolving community structure in complex networks," *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, pp. 7327–7331, 2007.

[13] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[14] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical review E*, vol. 76, no. 3, p. 036106, 2007.

[15] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the national academy of sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[16] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Physical review E*, vol. 68, no. 6, p. 065103, 2003.

[17] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[18] A. Lancichinetti, M. Kivelä, J. Saramäki, and S. Fortunato, "Characterizing the community structure of complex networks," *PloS one*, vol. 5, no. 8, p. e11976, 2010.

[19] H. Almeida, D. Guedes, W. Meira Jr, and M. J. Zaki, "Is there a best quality metric for graph clusters?" in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2011, pp. 44–59.

[20] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[21] E. A. Leicht and M. E. Newman, "Community structure in directed networks," *Physical review letters*, vol. 100, no. 11, p. 118703, 2008.

[22] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.

[23] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[24] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Physical review E*, vol. 78, no. 4, p. 046110, 2008.

[25] J. Shetty and J. Adibi, "The enron email dataset database schema and brief statistical report," *Information sciences institute technical report, University of Southern California*, vol. 4, 2004.