# Cosine Similarity for Multiplex Network Summarization

Athanasia Polychronopoulou
Temple University
Philadelphia, PA 19122 USA
Email: n.polychr@temple.edu

Fang Zhou
School of Data Science & Engineering
East China Normal University
Shanghai, China
Email: fzhou@dase.ecnu.edu.cn

Zoran Obradovic
Temple University
Philadelphia, PA 19122 USA
Email: zoran.obradovic@temple.edu

***Abstract*—Most of the natural systems encountered in all kinds of disciplines consist of a set of elementary units connected by relationships of different kinds. These complex systems are commonly described in terms of networks, where nodes represent the entities and links represent their interactions. As multiple types of distinct interactions are often observed, these systems are described as multiplex networks where the different types of interactions between the nodes constitute the different layers of the network. The ever-increasing size of these networks introduces new computational challenges and is therefore imperative to be able to eliminate the redundant or irrelevant edges of a network and create a summary that maintains the intrinsic properties of the original network, with respect to the overall structure of the system. In this work, we present a summarization technique for multiplex networks designed to maintain the structural characteristics of such complex systems by utilizing the intrinsic multiplex structure of the network and taking into consideration the inter-connectivity of the various graph layers. We validate our approach on real-world systems from different domains and show that our approach allows for the creation of more compact summaries, with minimum change of the structure evaluation measures, when compared to baseline methods that aggregate contributions of multiple types of interactions.**

## I. Introduction

Complex network theory has been well established as one of the main tools for understanding and analyzing the behavior of the natural systems that surround us. Most of these systems can be seen as a collection of entities interacting with each other and can be represented as complex networks, whose nodes (entities) are connected through edges (interactions). As network theory evolves, it becomes more apparent that these complex systems are commonly composed of multiple types of interactions, each carrying a different piece of information. For example, in social sciences, individuals may be connected by family, friendship, or professional ties, in biological sciences, proteins are commonly connected via different types of genetic and physical interactions. Such complex systems are represented in the form of multiplex networks, where each layer represents a different type of interaction among nodes. In addition to the interactions among the nodes of the networks, these systems also present correlations among the various types of interactions. These correlations are represented by the intrinsic structure of the network and are seen as associations of the various layers of the graph. For example, in social sciences, a network with a large overlap between two layers that represent two distinct types of people interactions i.e. friendship and professional ties might indicate that there is an interconnection between the two in the given network.

Mathematically, a Multiplex Network $G$ with $L$ layers can be seen as a collection of $L$ single-layer networks: $G = \{G_a | a \in \{1, ..., L\}\}$. Each of these networks has a set of edges, and they all share the same set of nodes. Then: $G_a = \{N, E_a, W_a\}$ , where $N$ is the common set of nodes, $E_a$ is the set of edges of layer $a$ or the intralayer connections of layer $a$, and $W_a$ is the set of weights of the corresponding edges, representing the connection strengths. In multiplex networks, the only possible type of interlayer connection is the one in which a given node is connected to its counterpart node in the remaining layers.

As the focus on the study of complex systems is continuously growing, graph summarization techniques become more crucial, offering significant benefits [1, 2]. These techniques can facilitate the observation of patterns otherwise hidden in the underlying data, by producing an overview of the social network that can be used for visualization. A summary of the graph can be handled easier and more efficiently and if the quality of the summarization is high, it carries most of the information on the original graph. Additionally, in cases of very large networks where query processing and data mining algorithms can be very inefficient, graph summarization, can enable the execution of complex analysis techniques [3].

There is no doubt that network summarization benefits the study of large graphs, eliminating some of the redundant information that they may carry. The question that arises is how to condense these graphs while retaining as much detail as possible about the whole system. For single-layer networks, a plethora of graph summarization techniques have been published, each using a different approach and having


ASONAM '21, November 8–11, 2021, Virtual Event, Netherlands © 2021 Association for Computing Machinery. ACM ISBN 978-1-4503-9128-3/21/11…$15.00 https://doi.org/10.1145/3487351.3488331

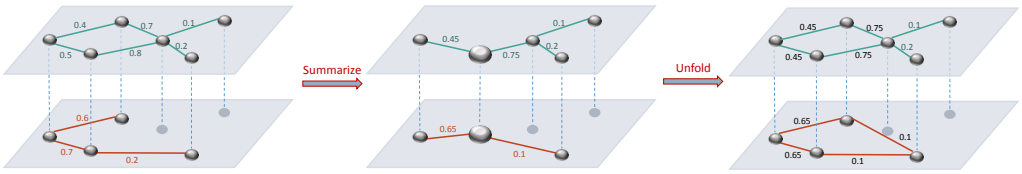


Fig. 1. Sample network summarization: The left side graph depicts a sample multiplex network of two layers that is summarized using the proposed approach. In the middle graph, the supernode represents all the nodes included in it and is connected via superedges to the rest of the graph. Each superedge represents all possible edges between the adjacent node and the nodes of the supernode. The right side graph is the multiplex network produced by unfolding the graph summary i.e. extracting the nodes from the supernode and assigning to them edges whose weights are determined by the weight of the superedge.

different sets of goals (i.e. compression, visualization, use with predictive algorithms) [1]. For multiplex networks, however, the summarization problem becomes more complicated, as the intrinsic structure of the layers carries additional information. So far, layers aggregation is used as a means of reducing the size of such networks [4]. This approach is most effective when some interaction layers are redundant or uninformative. However, if the various layers have minimum overlap, layer aggregation results in loss of information. The problem of the exact structural equivalence of graphs is also studied mainly by methods focused on identifying subgraphs within a network [5]. However, such methods are not easily implemented as network summarization techniques due to their high computational complexity. Another similar problem is that of graph clustering or partitioning [6] aiming to find collections of strongly related nodes, grouping them based on their direct connections. In contrast, graph compression aims at the reduction of the size of the graph and nodes are grouped based on the similarity of their relationships to other nodes.

In this work, we propose a method for complex network summarization focusing on multiplex networks of weighted and undirected networks, although, the method could be adjusted to work with directed networks as well. It is a summarization method based on iterative node-grouping, where nodes are grouped into supernodes, connected via superedges. While previously published methods may be extended for use with multiplex networks [7], these extensions aggregate or average the contributions of the different graph layers. On the contrary, our method aims to reduce the graph size by utilizing the intrinsic structure of the network and taking into consideration the inter-connectivity of the various graph layers when selecting the nodes to merged.

Our results indicate that this approach generates graph summaries of smaller size, while maintaining the informative content. In previous studies, the evaluation was either theoretical, such as finding the most economical description while avoiding maximum redundancy [8] or application and methodology specific [9, 10, 11]. In a different work, the graph summary was evaluated based on its distance from the original graph [7]. However, these approaches do not take into account the graph's structural properties that also need to be maintained. Therefore, we use a three-fold evaluation, utilizing a variety of well-established measures of graph comparison, applying data mining techniques to verify that important structural aspects are maintained and using a collection of structural descriptors that when merged can synthesize a very informative graph structure comparison. Finally, we test our procedure on real-world networks of different domains.

## II. Methods

The proposed summarization method for weighted and undirected multiplex networks implements an aggregation-based technique where similar nodes are grouped to supernodes and edges are grouped to superedges. A supernode then represents all the nodes included in it and a superedge represents all possible edges between all pairs of nodes in the adjacent supernodes. In the compressed graph, self-edges are also incorporated representing self-edges of nodes in the original graph or edges among different nodes in the supernode.

In every step of our iterative approach two nodes are first chosen for merging. Then both nodes are replaced by a new supernode, and their adjacent edges in each layer are aggregated and represented by superedges whose weights are calculated as the mean weight of all the edges they represent. An example of such a step is shown in Figure 1. Once this iteration is completed the process is repeated on a newly formed graph, to find the next pair of nodes to be merged (Algorithm 1).

**Algorithm 1**

```
procedure CREATE GRAPH SUMMARY
Input: G = {G_a | a ∈ 1, ..., L}
    while There are nodes to be merged do
        Choose the two nodes to be merged
        Merge the two nodes
        Update Mapping of nodes
        Update the Correction list
        Recalculate benefits of merging
    return
    end while
Output: For each step: SummarisedGraph,
            Node Mapping, Corrections List
```

A key concept of the method is the ability to choose in every step, the two nodes that have the most similar connectivity patterns. In a social sciences domain, for example, we would choose two people with multiple common acquaintances (high structural equivalence) and not people with similar attribute values. A mathematical measure that quantifies this idea, takes into account the weights of the connections, and can be directly extended to multiplex networks is the cosine similarity of the two nodes [12]. For two vectors $\mathbf{u}$ and $\mathbf{v}$ the cosine similarity is given by $\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$. For weighted networks, the

vectors $\mathbf{u}$ and $\mathbf{v}$ represent the connectivity vectors of nodes $u$ and $v$, including the weights of the links of each node with every other node in the network. For multiplex networks these vectors can be extended to be the concatenated connectivity vectors of all layers, i.e. $\mathbf{u}^\top = (\mathbf{u_1}^\top, \mathbf{u_2}^\top, ..., \mathbf{u_L}^\top)$. The use of the cosine similarity has several benefits. It guarantees that the selected nodes of every step have the largest percentage of common neighbors and the most similar weights among any other pair of nodes. It works well with multiplex weighed networks, and it is also not degree biased, treating nodes of high and low degree equally.

Utilizing the concept of structural equivalence of nodes, there are two possible implementations of the iterative approach. The first implementation is a greedy approach (G-CS method), where in every step the cosine similarity of all pairs of nodes is calculated so that the optimal pair is selected to be merged. The second implementation is a semi-random approach (SR-CS method), where in every step the first of the two nodes is randomly selected among the nodes of the graph and the second node is the best out of the first node's neighbors, i.e. the node that given the first selection provides the highest result for cosine similarity. This second approach does not guarantee that the nodes selected in every step are the optimal ones. However, it carries some of the properties of the proposed method (as it utilizes the cosine similarity as well), it is not degree biased and provides a significant improvement in the speed of the node selection algorithm.

After selecting the two nodes to be merged, the graph is updated so that it now includes the newly created supernodes and superedges. In each layer, the superedge weights correspond to the average weight of the layer's included edges (with missing edges being assigned a weight value zero). The remaining portion of the procedure focuses on producing the rest of the output of the algorithm, i.e. the detailed mapping of the nodes including the list of nodes in every supernode and a corrections list that could be used to reproduce the original graph if needed.

Finally, the values of similarity in connectivity patterns among nodes (and supernodes) need to be recalculated as a part of the graph has changed. The easiest approach would be to recalculate every single value (one value for each possible pair of nodes), however, this is a slow process, especially for very large networks. A more efficient approach is to recalculate the new similarity values only for the part of the graph that is affected. This part includes all the one and two-hop neighbors of the merged nodes, as well as nodes that were brought closer together because of the newly created node.

## III. Evaluation of Network summarization

### A. Baseline Methods

The value of our approach is demonstrated comparing the results with two baseline methods that follow the same iterative approach but differ in the node pair selection mechanism.

*1) Minimizing Graph Distance:* The first of the baselines is a natural extension to multiplex networks of the work published at [7]. The goal of that work was to produce a compressed graph that when unfolded produces a graph with the smallest possible distance from the original one. Extending their graph distance measure to multiplex networks we can calculate the distance of two graphs $G_a$ and $G_b$ as:

$$D_{a,b} = \sqrt{\sum_{l=1}^{L} \sum_{\{(u,v)\}_l \in V \times V} (w_a(u,v)_l - w_b(u,v)_l)^2} \quad (1)$$

In this equation $w_a(u,v)_l$ and $w_b(u,v)_l$ are the weighs of the edges between nodes $u$ and $v$ in layer $l$ of graphs $a$ and $b$, respectively. The first summation is over all layers $l$ and the second summation is over all pairs of nodes $u$ and $v$ that exist in the given layer $l$. Notice that in our setting graphs $a$ and $b$ would represent the original and the unfolded graph.

The distance-based summarization again selects nodes that have a large percentage of common neighbors and edges of very similar weights. However, this method is degree biased, showing a preference in low degree nodes whose small number of edges naturally produces a smaller distance in equation 1.

*2) Fully-Random node selection:* The second baseline method is based on a random selection of two neighboring nodes (FR method). This approach is not expected to maintain the structural characteristics of the graph as well, it does offer, however, an improvement in the speed of the node selection process. It is a method that can be considered in large scale applications or cases when fast summarization is required.

### B. Compression Evaluation Measures

The different network summarization techniques are first evaluated comparing the resulting graph summaries. Two major aspects are considered: the size of the final network, in terms of the number of edges after a given number of algorithmic iterations as well as the quality of the graph summary in terms of resemblance to the original network. The following detailed measures are utilized:

*1) Compression Ratio:* As compression ratio we define the ratio of the number of edges in the summarized graph with the number of edges in the original graph. It represents the reduction in the size of the network. Since the methods presented in this work reduce the number of nodes with a constant rate in each algorithmic iteration, the preferred method reduces the number of edges at a higher rate.

*2) Euclidean Distance from original graph:* This measure was introduced in [7] and is based on equation 1. Once normalised with the number of nodes, it quantifies the difference between the original graph and the graph that is unfolded from the summarized one. Naturally, the distance-based baseline provides a graph with minimal difference from the original one, as it is designed to provide just that. However, the greediness of this approach does not guarantee a global minimization of the distance.

*3) Graph Edit Distance from original graph:* The Graph Edit Distance (GED) is a well established measure of similarity between graphs that acts as a measure of topology change, and in this context it is studied in detail in [13]. For two graphs $G_a = (V_a, E_a)$ and $G_b = (V_b, E_b)$ GED evaluates

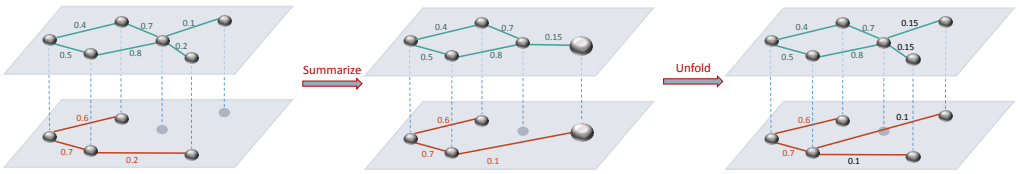


Fig. 2. Distance Baseline - Sample Network summarization: The left side graph depicts the same multiplex network of two layers as the previous figure. The summarized graph is now generated using the distance-based baseline. The corresponding unfolded graph is also shown on the right part of the figure.

the sequence of elementary graph edit operations required to modify an input graph $G_a$ such that it becomes isomorphic to a reference graph $G_b$ and is given by:

$$GED_{a,b} = \frac{|V_a| + |V_b| - 2|V_{ab}| + |E_a| + |E_b| - 2|E_{ab}|}{|V_a| + |V_b| + |E_a| + |E_b|} \quad (2)$$

where $V_{ab}$ and $E_{ab}$ indicate the node and edge overlap of the two graphs with $V_{ab} = V_a \cap V_b$ and $E_{ab} = E_a \cap E_b$

*4) Weighted Distance from original graph:* GED does not take into account the weighted nature of the networks. This property can be considered by summing the differences in edge-weight value over all edges in the two graphs. This measure was studied in [13] and can be defined as:

$$DEW_{a,b} = \sum_{u,v \in V} \frac{|w^a_{uv} - w^b_{uv}|}{max(w^a_{uv}, w^b_{uv})} \quad (3)$$

### C. Evaluation of Node Community Assignments

An important aspect of the original graph that characterizes the graph intrinsic structure and should be maintained by a graph summarization method is the node community assignments. The consistency of these community assignments is evaluated using an algorithm introduced in the work of [6], designed specifically for multiplex networks. In our setting, the community assignments in the original graph are compared against the community assignments in the unfolded graph.

### D. Structural Evaluation Measures

The different network summarization techniques are also evaluated comparing the intrinsic structural characteristics of the resulting graph summaries. For this purpose we utilize a variety of network descriptors.

*1) Intra-Graph Edge Overlap:* The edge overlap measure was originally introduced to evaluate the correlation of the node connectivity patterns between layers [14, 15]. For our purposes, this measure is redefined to compare the node connectivity patterns between the corresponding layer of the original and the unfolded graph and is given by

$$E^l_{G,G'} = \frac{1}{|E_P|} \sum_{i,j} A^l_{ij} \cdot U^l_{ij} \quad (4)$$

where $A^l_{ij}$ is the adjacency matrix of layer $l$ in the original graph $G$ and $U^l_{ij}$ is the adjacency matrix of layer $l$ in the unfolded graph $G'$. In this context, $E^l_{G,G'}$ is a normalized measure of the number of edges that $A^l_{ij}$ and $U^l_{ij}$ have in common. The normalization constant $|E_P|$ represents the number of links that the two graphs could have in common, i.e. the number of edges in the network projected by the two graphs. Finally, the intra-graph edge overlap is calculated as the average of the edge overlap of all the layers: $E_{G,G'} = mean(E^l_G)$.

*2) Degree distribution:* Another characteristic of a graph that can offer insight into the structure of each layer is the distribution of the nodal degree [16]. In order to compare the degree distribution of each layer of the original graph with the degree distribution of the corresponding layer in the summarized graph, we use a measure that is commonly used in statistics for the comparison of probability distributions, the Jehnsen-Shannon Divergence. This is a symmetric measure with values between 0 and 1, having therefore all the characteristics of a distance measure. Notice that, the Jehnsen-Shannon Divergence of the two graphs is calculated as the average of the values of all the layers.

*3) Layer Pairwise Multiplexity:* In most multiplex networks not all nodes have connections in all layers. Then, every pair of layers may or may not contain the same nodes and may or may not have correlated node activity patterns. The measure of Layer Pairwise Multiplexity (LPM) introduced in [17] quantifies this correlation between layers $a$ and $b$:

$$Q_{ab} = \frac{1}{N} \sum_{i=1}^{N} \beta_{i,a} \beta_{i,b} \quad (5)$$

where $\beta_{i,a}$ is the activity of node $i$ on layer $a$ and takes the values 1 and 0 depending on whether the node is active or not on the layer. For our purposes, this measure can be used as an evaluation of the degree at which each summarization method is changing the activity pattern correlations of the layers. Such an example is evident by comparing Figures 1 and 2, where it can be seen that the distance-based method generates an unfolded graph with a node that was not present in the original graph, changing the LPM of the two layers. Since the measure characterizes pairs of layers, the average LPM between all pairs of layers is calculated for the original graph and then compared with the corresponding LPM calculated on the unfolded graph.

*4) Graph Clustering Coefficient:* The graph clustering coefficient was originally introduced by Watts and Strogatz [18] and it quantifies the tendency of nodes to form triangles, following the popular saying "the friend of my friend is my friend". It's extension for use with multiplex networks [19] considers triangles that are formed not only using the intralayer links, but also the interlayer links of nodes with their counterparts. An example of such a natural situation arises

| | FAO Data | | | | BTS Data | | | | *C. elegans* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Edges | G-CS | SR-CS | Dist. | FR | G-CS | SR-CS | Dist. | FR | G-CS | SR-CS | Dist. | FR |
| 95% | 9 | 6 | 42 | 12 | 70 | 14 | 78 | 22 | 353 | 282 | 354 | 351 |
| 90% | 14 | 13 | 53 | 26 | 95 | 21 | 134 | 49 | 718 | 588 | 718 | 686 |
| 70% | 27 | 38 | 84 | 64 | 144 | 96 | 211 | 143 | 1781 | 1653 | 2121 | 1960 |
| 50% | 49 | 66 | 105 | 95 | 184 | 177 | 254 | 209 | 2666 | 2583 | 2829 | 2999 |

TABLE I

THE NUMBER OF ALGORITHMIC ITERATIONS REQUIRED TO SUMMARIZE THE ORIGINAL NETWORK TO THE LEVEL WHERE THE NUMBER OF EDGES EQUALS THE GIVEN PERCENTAGE OF THE NUMBER OF EDGES IN THE ORIGINAL NETWORK.

from a social sciences network where one person $i$ knows $j$ from the university, $i$ also knows $k$ from work while $j$ and $k$ know each other from a reading club. In these cases the clustering coefficient of node $i$ in a multiplex network is given by

$$C(i) = \frac{2 \sum_{\ell=1}^{L} |E_\ell(i)|}{\sum_{\ell=1}^{L} |N_\ell(i)|(|N_\ell(i)| - 1)} \tag{6}$$

where $|E_\ell(i)|$ is the number of edges in the subgraph generated by all the neighbors $N(i)$ of node $i$ (regardless of the layer in which there is an edge) and the edges among these nodes on layer $\ell$. Also $N_\ell(i)$ represents the subset of $N(i)$ that is active on layer $\ell$. Then the clustering coefficient of the network $G$ can be defined as the average of all the $C(i)$.

The clustering coefficient presented in Equation 6 suffers from a major limitation, its outcome does not take into consideration the weight of the edges in the network. To overcome this limitation we evaluate a different measure that was originally introduced by [20] for weighted single-layer networks. This measure takes into account the importance of the clustered structure by measuring the interaction intensity of the local triangles and is now extended for use with multiplex networks. For node $i$ we calculate:

$$C_w(i) = \frac{\sum_{\ell=1}^{L} \sum_{j,h} \frac{(w_{ij}^\ell + w_{ih}^\ell)}{2} a_{ij}^\ell a_{ih}^\ell a_{jh}^\ell}{\sum_{\ell=1}^{L} s_i^\ell (k_i^\ell - 1)} \tag{7}$$

where $w_{ij}^\ell$ is the weight of the edge between nodes $i$ and $j$ on layer $\ell$, $a_{ij}^\ell$ is the entry of the adjacency matrix corresponding to the edge between nodes $i$ and $j$ on layer $\ell$, $k_i^\ell$ is the degree of node $i$ on layer $\ell$ and $s_i^\ell$ is the strength of node $i$ on layer $\ell$ defined by $s_i^\ell = \sum_{j=1}^{N} a_{ij}^\ell w_{ij}^\ell$. The clustering coefficient of the network $G$ can be defined as the average of all the $C_w(i)$.

### *E. Data*

We evaluate our method on three real-world datasets. First, we consider the network obtained using the food and agricultural trade data from the Food and Agriculture Organization of the United Nations [21] (FAO Data). In this network, nodes represent countries connected using a symmetric measure of Import-Export quantities of specific products between the two countries. We consider five different products (Dried Fruit, Macaroni, Margarine, Chicken, Prepared Nuts) creating a multiplex network of five layers and 299 nodes.

The second dataset comes from the Bureau of Transportation Statistics of the United States Department of Transportation [22] (BTS Data). The data comprise the domestic segment of the database and describe all the flights performed by all carriers in 2014. In this network the 1222 US airports represent nodes and the edges are created by a symmetric and normalized measure of the number of departures between the two airports. Finally, each of the five layers of the multiplex network corresponds to one of the five largest airlines (the size is measured by the total number of departures).

The third dataset is comprised of the complete set of protein-protein interactions of *C. elegans*, as obtained from BioGRID [23]. This is an un-weighted multiplex network of 3 layers (3 types of interactions) formed by more than 3500 nodes (proteins) and 10000 edges (protein-protein interactions).

## IV. RESULTS

Results are presented separately for the three types of evaluation measures introduced in Sections III-B - III-D: compression evaluation, node community assignments and graph intrinsic structure. Our experiments aim to address several questions: 'How efficient is each of the methods in networks compression: how many iterations of the algorithms are required?', 'How much is the network changing as it is being summarized?', 'How does compression affect the intrinsic clustering of nodes?', and 'How much can we summarize the network, while maintaining its structural characteristic?'.

The experimental setup is such that on each of the datasets all four summarization methods are applied: both implementations of the proposed method that incorporates the cosine similarity (Greedy and Semi-Random) and the two baseline methods (Minimizing Graph Distance and Fully-Random node selection). For each method, and after every summarization step, the unfolded graph is generated and all the evaluation measures are calculated.

***Compression Ratio:*** 'How efficient is each of the methods in networks compression?'. This question may be answered using the results presented in Table I. The table lists the number of algorithmic steps required to create a network summary with the given percentage of edges. In all data sets both implementations of the proposed method reduce the number of edges at a faster rate than the baseline methods. For example, to reduce the number of edges to 90% of that of the original network in FAO data, the proposed methods

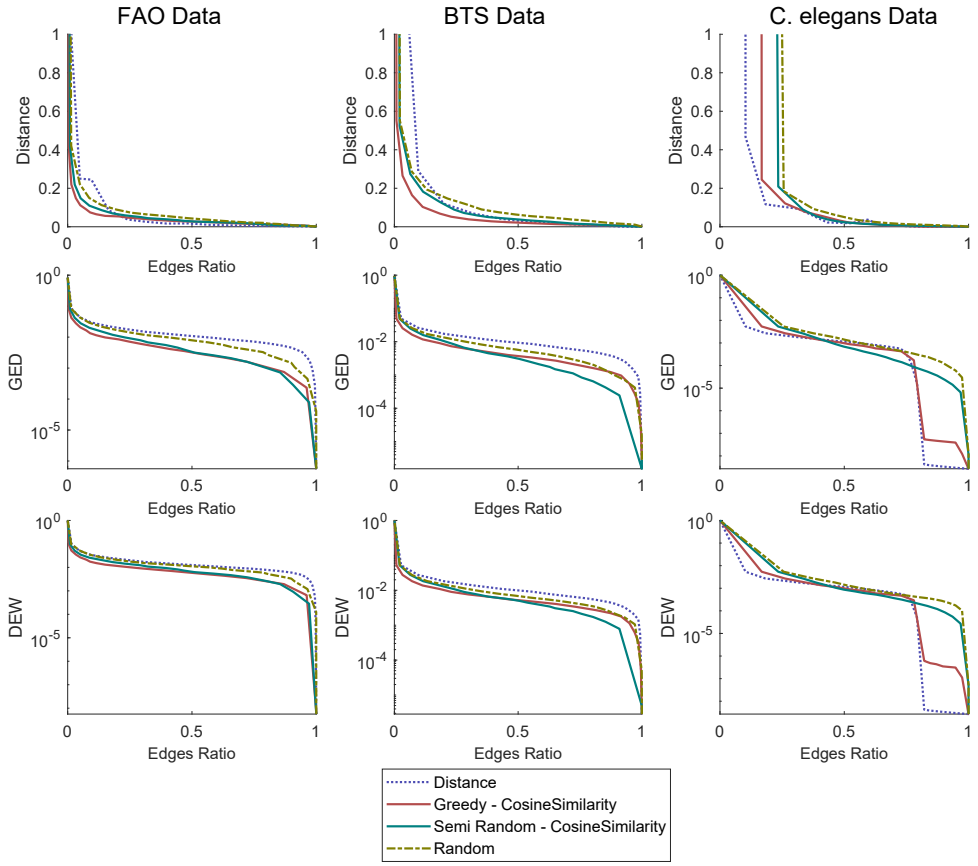


Fig. 3. The evolution of the three distances between the original and the unfolded network, as generated after each of the summarization steps. The x axis represents the Edges Ratio, whose value, as the summarization progresses and the number of edges becomes smaller, drops from 1 towards 0.

require about one-quarter of the steps needed by the Distan minimizing method. Notice that, the smaller the number iterations required, the more efficient the process is and in t case of *C. elegans*, which also represents the larger of o datasets, the difference is mostly observed at higher summ rization percentages and offers a significant time reduction.

***Distance from the original graph:*** 'How much is t network changing as it is being summarized?'. To answer th question, we use the three distance measures introduced befc and plot their evolution as the network is being summarize i.e. as the edges ratio in the summary drops from one towar zero. The results, presented in Figure 3, indicate that bc implementations of our approach perform better than t baselines, since the graphs are changing at a slower pace. the case of graph distance, one of the baselines was design explicitly to create a summary that is the least different frc the original network and therefore it is expected to outperfor the rest of the methods. However, for the FAO data, when the graph was compressed to half of the original size, all methods, except the random approach, generated graphs that were equally different from the original one. On the BTS data, the Greedy Implementation of our proposed method even outperforms the baseline. This fact along with the non-smooth line of the distance-based plot is explained by the greediness of the nature of the Distance minimizing baseline method.

***Node Community Assignments:*** 'How does compression

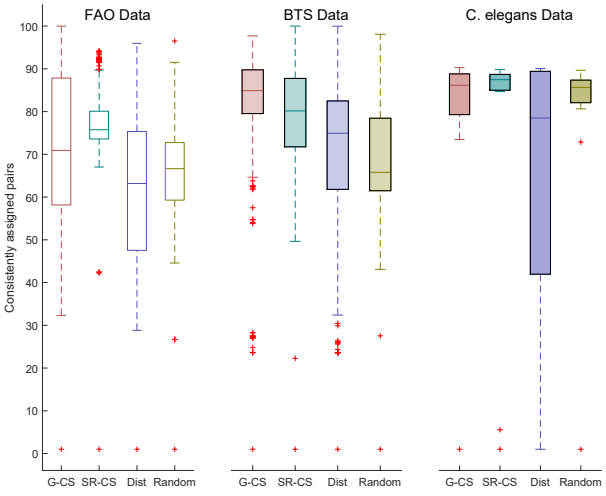


Fig. 4. The percentage of node pairs that have consistent clustering assignments in the unfolded and the original graph (i.e. node pairs that consistently belong to the same or a different cluster).

affect the intrinsic clustering of nodes?'. To answer this question we study the consistency of the assignment of nodes in communities for each of the summarization methods. After each summarization step the unfolded network is generated and the community assignment algorithm is executed. Then, we calculate the percentage of pairs of nodes that have been assigned in the same or a different cluster, when compared

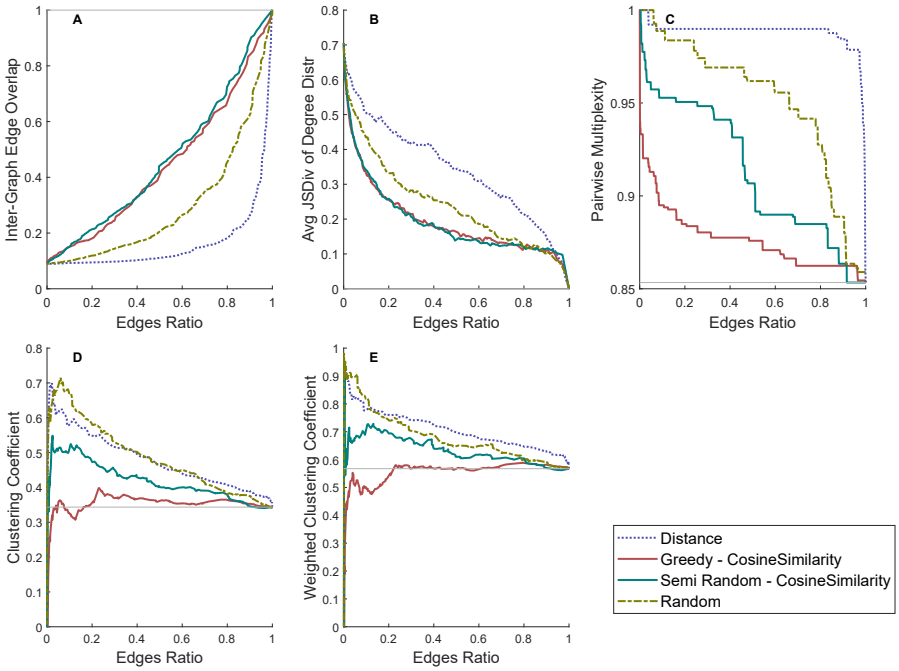


Fig. 5. Comparison of the results for the two methods proposed and the two baseline methods for the FAO database. The various evaluation measures are plotted against the edge ratio which represents the degree of summarization of the graph. In all plots, the gray line represents the starting value, i.e. the value of the measure at the original network. As the summarization progresses and the number of edges in the summarized network becomes smaller the value of edge ratio drops from 1 towards 0.

with their cluster assignment in the original graph. The results, that include all the summarization steps, are presented in Figure 4. For all three data sets, it is evident that the proposed approach maintains more successfully and more consistently the natural clustering of the original graph.

***Structural Characteristics:*** 'How much can we summarize the network, while maintaining its structural characteristic?'. To answer this question we present in Table II the maximum possible edge compression ratio under the restriction that the given structural measures need to remain within the specified percentage of their initial value. Then, for example using the proposed method, in order to remain within 90% of the original value of the Intra-Graph Edge Overlap, the network can be summarized to the level where the number of edges becomes 93% of their original number. If instead, we use the minimum distance method or the fully random one we can only achieve a summary of about 99% of the original edges. Similar observations can be made for all the measures, as indicated by the bold values of Table II marking the best performing algorithm in each case. It is also worth noticing that some characteristics are much more robust to summarization than others. For example, we can use the proposed method to summarize the FAO dataset network to the level of about 16% of the original edges (edge compression ratio 0.163) and the value of the weighted clustering coefficient will still be within 90% of its original value. This is not however the case for the two baseline techniques, for which the edge ratio can only get to 0.917 for the distance minimizing method and 0.726 for the random approach. More detailed results are plotted in Figure 5 in order to answer the more general question: 'How does compression affect the intrinsic clustering of nodes?'.In these plots the evolution of the value of each of the measures is presented as the summarization progresses from an initial value of edge ratio 1 towards the final value 0. The robustness of the clustering coefficients becomes now more evident, as it remains practically stable until extreme values of edges ratio. It also becomes clear that all the measures are changing at a slower rate when the proposed summarization methods are used, creating this way network summaries that are more representative of the original network.

| | | FAO Data | | | |
|---|---|---|---|---|---|
| | | G-CS | SR-CS | Dist. | FR |
| Intra-Graph Edge Overlap | 95% | **0.982** | **0.981** | 0.996 | 0.989 |
| | 90% | **0.936** | **0.933** | 0.994 | 0.967 |
| | 70% | **0.825** | **0.823** | 0.982 | 0.918 |
| | 50% | **0.660** | **0.660** | 0.964 | 0.848 |
| Degree JSDiv | 95% | **0.982** | 0.999 | **0.982** | 0.984 |
| | 90% | **0.965** | 0.999 | 0.968 | **0.960** |
| | 70% | **0.306** | **0.302** | 0.803 | 0.510 |
| | 50% | **0.083** | **0.086** | 0.464 | 0.172 |
| P. Multiplexity | 95% | 0.962 | **0.916** | 0.998 | 0.966 |
| | 90% | **0.637** | 0.880 | 0.998 | 0.913 |
| | 70% | **0.083** | 0.511 | 0.997 | 0.848 |
| | 50% | **0.014** | 0.456 | 0.994 | 0.806 |
| Clust. Coeff. | 95% | **0.795** | 0.859 | 0.993 | 0.918 |
| | 90% | **0.268** | 0.818 | 0.917 | 0.818 |
| | 70% | **0.008** | 0.302 | 0.525 | 0.546 |
| | 50% | **0.005** | 0.026 | 0.262 | 0.296 |
| W. Clust. Coeff. | 95% | **0.191** | 0.832 | 0.988 | 0.860 |
| | 90% | **0.163** | 0.487 | 0.917 | 0.726 |
| | 70% | 0.008 | **0.004** | 0.351 | 0.240 |
| | 50% | **0.004** | **0.004** | 0.025 | 0.070 |

TABLE II

THE MAXIMUM ACHIEVABLE EDGE COMPRESSION RATIO WHEN CREATING A GRAPH SUMMARY THAT MAINTAINS THE STRUCTURAL MEASURES WITHIN A GIVEN PERCENTAGE OF IT'S INITIAL VALUE.

## V. Conclusion

Motivated by the recent increase in the use of Multiplex networks, whose exploration and utilization become increasingly difficult, we propose a network summarization approach for weighted multiplex networks. Our method focuses on removing structural redundancy while maintaining the information carried by the intrinsic structure of the graph. Using real-world data from different domains, our method is shown to maintain more accurately the properties of the original graph and for a larger summarization percentage. Conversely, the distance-based approach and even more so the random approach significantly alter the graph characteristics, leading to a graph summary that should not be used as a guide for the description, optimization, or calculation of statistics of the original network. Furthermore, the proposed method is shown to reduce the size of the network, as this is represented by the number of edges, faster than the baselines resulting in a more efficient summarization technique. Finally, the greedy implementation of the proposed method is computationally comparable to the distance-based approach as the bottleneck in both cases is the search of the optimum pair of nodes for each step. However, our results indicate that the semi-random approach carries most of the benefits of the greedy implementation and is additionally computationally competitive, reducing significantly the time required for each step, and allowing the method to be applied to larger datasets. Notice that, the presented results are restricted to smaller size networks simply because the calculation of the evaluation measures is not efficient. The increased complexity of the calculations for the clustering coefficient, pairwise multiplexity, and communities' detection restricted the size of the networks used for the evaluation and presentation of results. The summarization algorithm itself can be applied networks of any size.

## References


[1] Yike Liu et al. "Graph summarization methods and applications: A survey". In: *ACM Computing Surveys (CSUR)* 51.3 (2018), pp. 1–34.

[2] Roberto Interdonato et al. "Multilayer network simplification: approaches, models and methods". In: *Computer Science Review* 36 (2020), p. 100246.

[3] Fang Zhou, Mohamed Ghalwash, and Zoran Obradovic. "A fast structured regression for large networks". In: *2016 IEEE International Conference on Big Data (Big Data)*. IEEE. 2016, pp. 106–115.

[4] Manlio De Domenico et al. "Structural reducibility of multilayer networks". In: *Nature communications* 6.1 (2015), pp. 1–9.

[5] Ben D MacArthur, Rubén J Sánchez-García, and James W Anderson. "Symmetry in complex networks". In: *Discrete Applied Mathematics* 156.18 (2008), pp. 3525–3531.

[6] Peter J Mucha et al. "Community structure in time-dependent, multiscale, and multiplex networks". In: *science* 328.5980 (2010), pp. 876–878.

[7] Hannu Toivonen et al. "Compression of weighted graphs". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011, pp. 965–973.

[8] Saket Navlakha, Rajeev Rastogi, and Nisheeth Shrivastava. "Graph summarization with bounded error". In: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 2008, pp. 419–432.

[9] Gregory Buehrer and Kumar Chellapilla. "A scalable pattern mining approach to web graph compression with communities". In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*. 2008, pp. 95–106.

[10] Yuanyuan Tian, Richard A Hankins, and Jignesh M Patel. "Efficient aggregation for graph summarization". In: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 2008, pp. 567–580.

[11] Ning Zhang, Yuanyuan Tian, and Jignesh M Patel. "Discovery-driven graph summarization". In: *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*. IEEE. 2010, pp. 880–891.

[12] M. Newman. "Networks: an introduction". In: *Oxford University Press* (2010).

[13] Horst Bunke et al. *A graph-theoretic approach to enterprise network dynamics*. Vol. 24. Springer Science and Business Media, 2007.

[14] Federico Battiston, Vincenzo Nicosia, and Vito Latora. "Structural measures for multiplex networks". In: *Physical Review E* 89.3 (2014), p. 032804.

[15] Ginestra Bianconi. "Statistical mechanics of multiplex networks: Entropy and overlap". In: *Physical Review E* 87.6 (2013), p. 062806.

[16] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks". In: *Reviews of modern physics* 74.1 (2002), p. 47.

[17] Vincenzo Nicosia and Vito Latora. "Measuring and modeling correlations in multiplex networks". In: *Physical Review E* 92.3 (2015), p. 032805.

[18] Duncan J Watts and Steven H Strogatz. "Collective dynamics of 'small-world'networks". In: *nature* 393.6684 (1998), pp. 440–442.

[19] Regino Criado et al. "A mathematical model for networks with structures in the mesoscale". In: *International Journal of Computer Mathematics* 89.3 (2012), pp. 291–309.

[20] Alain Barrat et al. "The architecture of complex weighted networks". In: *Proceedings of the national academy of sciences* 101.11 (2004), pp. 3747–3752.

[21] *UN Food and Agriculture Organization*. http://www.fao.org. 2019.

[22] *U.S. Department of Transportation (USDOT), "Research and Innovative Technology Administration." (RITA)*. 2019.

[23] Chris Stark et al. "BioGRID: a general repository for interaction datasets". In: *Nucleic acids research* 34.suppl1 (2006), pp. D535–D539.