An Empirical Study of Network Reduction: The Measurement and Comparison

Chao-Lung Yang

clyang@mail.ntust.edu.tw

National Taiwan University of Science and Technology

Ming-Chieh Cheng

National Taiwan University of Science and Technology

Apicha Lumveerakul

National Taiwan University of Science and Technology

Yu-Wei Hsu

National Taiwan University of Science and Technology

Po-Sen Lai

National Taiwan University of Science and Technology

Research Article

Keywords: social network analysis, community detection, network reduction, network measurement, normalized adjusted ratio sampling

Posted Date: October 5th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3394930/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Social Network Analysis and Mining on June 14th, 2024. See the published version at https://doi.org/10.1007/s13278-024-01276-z.

An Empirical Study of Network Reduction: The Measurement and Comparison

Chao-Lung Yang*, Ming-Chieh Cheng, Apicha Lumveerakul, Yu-Wei-Hsu, and Po-Sen Lai

Department of Industrial Management, National Taiwan University of Science and Technology, Taipei, Taiwan

*corresponding author: Chao-Lung Yang, clyang@mail.ntust.edu.tw

ABSTRACT

Large-scale networks face challenges for analysis and visualization in social network analysis due to their enormous size. Network reduction and clustering are essential techniques for large-scale networks. This study proposed an analytic framework that combines degree distribution, clustering coefficient distribution, KS-statistic, and normalized adjusted ratio sampling (NARS) to measure the social network dataset before and after reduction. The proposed NARS ensures that the network can obtain a fair share of nodes based on cluster size. The proposed framework aims to compare and investigate the effectiveness of network reduction and clustering. To evaluate the framework, 20 datasets of undirected networks were tested. Results show that the proposed framework is able to compare the reduced network to the original network. Based on the experimental results, random walk, one of the network reduction methods, and its improved version, induced subgraph random walk methods, perform equivalently although random walk can provide faster computational time.

Keywords: social network analysis, community detection, network reduction, network measurement, normalized adjusted ratio sampling.

1. Introduction

Social network research is a field of social science that studies the social relationships and the interactions of individuals, focusing on quantifying and analyzing the structure, organization, activities, and behaviors of human social systems [1]. Social network research includes several domains: 1) the transmission of knowledge or information through various mediums, connections to certain objects, or communication interactions [2], 2) the prediction of relationships among social entities[3], and 3) the analysis of the impact of social phenomena [4].

The social network involves a set of nodes (or network members) that are linked by one or more types of relationships [5]. Most of these nodes or vertices are persons or organizations, but everything can be considered as a node in principle. For example, web pages [6], journal articles [7], countries [8], neighborhoods, or departments within organizations [9]. Social network analysis (SNA) combines the concepts from the structure and topology, branches of mathematics and graph theory, to study behaviors in the social networks [10]. SNA can be used to explain various social phenomena. Back in 1952, psychiatrist Moreno used "sociometry" to graphically represent how individuals feel about each other [11]. Sociometry provides ways to quantify abstract social concepts. Harary and Norman also used graph theory for the first time to analyze social networks [12].

Social networks can be complex and include millions, tens of millions, or even more nodes and edges, making them difficult to visualize and understand [13]. Such complexity can also result in a significant waste of computing power [14]. In the literature, there are two research areas: network reduction and community detection aiming to reduce the network size while preserving topology and node properties by

simplifying the network and minimizing information loss to obtain the most representative network [15].

Essentially, network reduction research includes four dimensions of the problems to be overcome: 1) structure preservation, 2) content preservation, 3) data sparsity, and 4) scalability [15]. When the scale of the social network is large, reducing the social network is usually costly in terms of time and computation power. Optimizing the algorithm with limited resources is the main challenge in studying network reduction in social networks. Various studies have been proposed to conduct topology-based computation or data mining techniques for reducing the network [16, 17]. Bhaumik proposed a graph-based approach to reduce community network graphs and associate sensors with meaningful communities by reducing the cost of analyzing and processing large amounts of sensor data [18]. These studies fall into the categories of "graph summarization," "graph reduction," or "graph representation," which can simplify the network graph.

Figure 1 illustrates the snapshots of the complete social network on the left and the graph in reduced search space on the right [18].

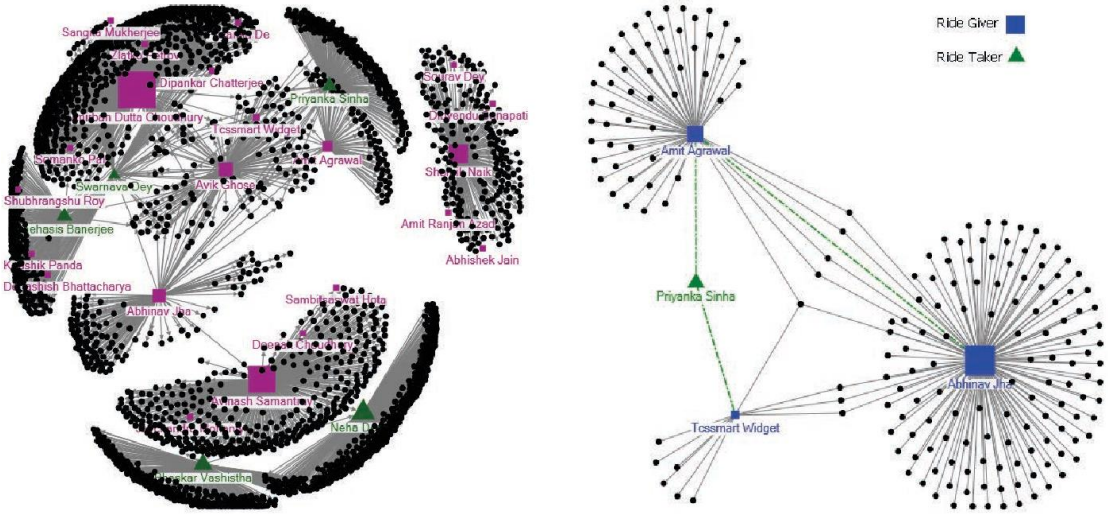

Figure 1 Snapshot of full social network and graph in reduced search space [18].

On the other hand, community detection is the method of network clustering, also

known as network clustering or grouping, that can identify the structure or clusters within a network. The real-world networks exhibit characteristic patterns such as power-law distribution in degree distribution [19, 20] or the "small-world phenomenon" [21-23], that represents clustering caused by short average distances between nodes and uneven connection distributions, also known as social structure. Based on the knowledge of the identified social clusters or structures, the information among networks can be extracted without using all of network.

Although, both network reduction and community detection tried to reduce the complexity of social networks, their methodologies are two distinct extensions of SNA. In this research, a data analysis framework was proposed to measure the effectiveness of combining network reduction and community detection when simplifying the social network with structure preservation. By comparing the efficiency of network reduction methods with and without community detection applied, this work offers a definitive way to quantifying social network datasets and evaluated if the combination of two methods can achieve better performance than the network reduction method only.

In this study, a framework to measure and compare the social network dataset before and after applying reduction method was proposed. To ensure the fairness of the comparison, the Normalized Adjusted Ratio Sampling (NARS) method was proposed to sample data nodes in each clustering group based on the size of the network. It uses a formula that considers the size of the whole network and the sizes of the groups. If a group is very small, it might not be included in the final selection. To evaluate the information preservation of the reduced network compared with the original one, two indicators, based on the literature, 1) degree distribution (DD) [24] and 2) clustering coefficient (CC) [25] were used to measure the number of the connections among nodes, and the number of nodes in the same cluster and identify small interlinked core areas

on a network, respectively. Then, KS distance [26] was used to calculate differences between complete and reduced networks based on the aforementioned DD and CC to evaluate the difference after reduction. In this work, 20 well-known public datasets of social networks with different sizes, structures, and features were used for conducting experiments.

The rest of this paper is organized as the following. Section 2 discusses the previous work, Section 3 delivers the methodology, Section 4 presents the experimental results, and Section 5 concludes this study.

2. Literature Review

2.1. Network Reduction

As the world's knowledge of network structure grows, complex network analysis with a large number of nodes and complicated connection structures becomes significant. The reasons for executing network reduction are primarily two-fold: the limitations of computational resources [35] and the difficulties in visualization understanding [13]. The limitation of computational resources is also a barrier to connecting network analysis and real-world applications, and good visualization can aid in understanding network structure and gaining insights from it.

According to Martin et al. study, the tasks of network reduction can be classified into three categories: partitioning, compression, and simplification [36]. There are several tools that have been applied to network reduction not only in a wide range of applications, but also in great variety of algorithms. Dubois and Bothorel proposed a network reduction method based on edge removing in 2005 which can be used to preserve important network attributes and structure [37]. Kudelka et al. proposed a network reduction method based on the “forgetting curve” in 2010 to simulate the

human brain learning and forgetting, and to apply its concepts to partition and reduce non-representative structures [38]. Oh proposed a network reduction method which is based on the congestion status of buses in 2012 based on a power-based factor named Power Transfer Distribution Factor (PTDF) to compute an appropriate bus grouping method to create network congestion [14].

Within those network reduction algorithms, random walk (RW) algorithm caught the attention in the literature for information extraction and network reduction. RW is a mathematical concept that models a path composed of random steps taken in different directions. In the context of a graph G, a RW involves starting from a node v_0, and then, at each step t , moving to a neighbor of v_t at the t step with probability $P_t = \frac{1}{deg(v_t)}$. where $deg(v_t)$ is the degree of node v_t, representing the number of edges connected to it. This process generates a sequence of random nodes. To capture the visited nodes and edges, a Markov chain can be constructed. However, if the RW falls into a loop, it might not explore enough nodes to meet the required sample size. To avoid this situation, another node can be selected, and the procedure repeated. In our experiment, we set the "long run" step as 100 * n, where n is the required number of nodes to ensure sufficient exploration [59]. In Tsugawa's and Ohsaki's research, a RW method was applied to the network graph, and then nodes that were traversed during the random walk were collected to build subgraphs [37]. This approach allowed for the extraction of localized structures or communities within the larger network, providing insights into the organization and connectivity patterns of the network at a finer scale.

Similarly, Kleinberg's study implemented RW with the induced subgraph called Induced Subgraph Random Walk (ISRW) to study the small-world phenomenon [38]. ISRW method introduces the concept of an induced subgraph, which is a subset of nodes and all the edges between them. The concept of induced subgraph came from

graph theory. Let G = (V, E) be any graph, and let S⊂V be any subset of vertices (nodes) of G. Then the induced subgraph G[S] is the graph whose vertex set is S and whose edge set consist of all of the edges in E that have both endpoints in S [60]. ISRW process is similar to RW, but with an additional step called "Graph Induction" to transform the subgraph into an induced subgraph. During this ISRW step, any additional edges that exist between the selected nodes are included. However, it is possible to fall into a loop of duplicated paths when running this method, and 100 * n steps are used as the limitation. In ISRW, very large graphs can be traversed in a relatively short number of steps in a graph by using random walk techniques.

Ahmed et al. proposed a novel graph sampling algorithm called Total Induced Edges Sampling (TIES) that aims to offset the bias introduced by random node and edge sampling [39]. Instead of randomly selecting nodes or edges, TIES leverages the concept of information entropy to identify critical nodes and edges that carry the most information about the graph's structure. TIES can be considered an edge-based method that utilizes the induced method, which involves adding other edges between the set of sample nodes. The main difference between the TIES method and simple edge sampling lies in the process of graph induction [39]. In TIES, the edges selected during the edge sampling step are augmented by adding other edges between the set of sample nodes. This process is known as graph induction. To perform graph induction, edges are initially selected randomly, and then any other edges that exist in the original graph between any of these sampled nodes are included. TIES starts by randomly selecting edges and subsequently adds any other edges that exist in the original graph between any of these sampled nodes. The steps in the TIES method can be summarized as follows: 1) node selection, and 2) graph induction

Following the aforementioned literature, this research applied the mentioned

network reduction methods such as RW, ISRW, and TIES as the basis methods of network reduction framework.

2.2. Community Detection

Community detection, also known as network clustering or grouping, aims to identify the structure or clusters within a network or graph composed of nodes and edges. It can greatly reduce the complexity of the original network graph and provide macro-level structural information and knowledge [40]. Real-world networks exhibit characteristic patterns such as power-law distribution in degree distribution [19, 20] or the "small-world phenomenon" [21-23], that represents clustering caused by short average distances between nodes and uneven connection distributions, also known as social structure.

Community detection or network clustering has extended to many fields. However, in social network analysis, network clustering refers to grouping nodes into clusters according to their characteristics and attributes. The simplest and most common definition is that a cluster corresponds to a set of nodes, and there are more connections within the cluster and fewer connections between the clusters and rest of the graph [45].

In the literature, multiple clustering methods were proposed. For example, in 2009, Fortunato investigated the community detection of non-overlapping communities and classified the techniques based on the principles of the method [46], In 2011, Coscia et al. explored overlapping community detection and classified the methods according to different definitions of community [47]. Crampes and Plantié also proposed an alternative classification method based on input and output data [48].

Among the many community models and algorithms, the most widely known method is the iterative two-stage algorithm designed by Blondel et al. [49] also known as Louvain method. This method uses modularity as a measure of connection density

within a community [62] The method consists of two main steps. First, each node is initially placed in its own separate community. Then, the algorithm goes through each node and tries to move it to a different community if doing so would maximize the modularity $\boldsymbol{Q}$ of the network. The modularity gain $\Delta\boldsymbol{Q}$ derived from adding node *i* to community C. The process repeats until no more improvements can be made. The goal is to find the best community arrangement that maximizes the modularity of the network. The second step is to construct a new network consisting of nodes that are those communities previously found. The process repeats until no more improvements can be made. The goal is to find the best community arrangement that maximizes the network's modularity. This algorithm reduces the time complexity of the Girvan-Newman algorithm to linearity, and its result also shows excellent results for network clustering.

However, study in 2019 by Traag et al. pointed out that the Louvain method may have some problems [50]. For instance, the resolution limit of the network may affect the number of communities discriminated causing specific flaws in the structure. Therefore, Traag et al. introduced a new method called Leiden which tried to improve randomization methods for better performance. Traag proposed an algorithm named the Leiden method, which ensures well-connected communities. Compared to Louvain's method, Leiden's algorithm consists of three steps: 1) This step is similar to Louvain's algorithm, where each node is added to a community to improve modularity. However, an important distinction is that Louvain's method searches all adjacent nodes, whereas Leiden's method optimally adjusts its search using heuristics [84]. This optimization focuses on nodes that can change neighboring communities. 2) Refinement of the partition. In this step the algorithm computes a new community structure adjusted from the original one. All nodes in the community are initially treated as independent, and

their ΔQ values are calculated to cluster the nodes within the community. Unlike the Louvain's method, this step only searches for the node within the community. 3) Aggregation of the network based on the refined partition. In this step the clusters found in the previous step are aggregated to create a new network for the next stage of computation. Their work demonstrated that the improved algorithm can provide explicit guarantees and bounds, as well as generate partitions that are stably connected within all communities. Since Louvain and Leiden methods have not been applied with the network reduction methods combined, this work evaluated both Louvain and Leiden methods with network reduction methods to check if there is a significant difference.

3. METHODOLOGY

3.1. Comparison of Original and Reduced Network

As mentioned before, analyzing social networks can be challenging due to their enormous size. Measuring graph/network similarity is one of major social network analysis to study graphs/networks. Essentially, graph/network similarity can compare the graphs/networks based on the selected characteristics or features that can be used to represent the graphs/networks. The graph similarity calculations can be classified into two categories: 1) graph theory indicators, including topological features of nodes or connection combinations, spatial location, and spectrum, and 2) algorithms based on image distance, such as Graph Edit Distance or Hamming distance[51]. In addition, Dubois proposed an algorithm that utilizes the shortest path and transitive closure to measure transitive properties [52]. Bai et al. research, extended from the field of neural networks, Graph Edit Distance (GED) is used as an indicator to measure network similarity and the quality function of the method [53].

In the literature, there are multiple ways to describe a network or graph in addition to the above methods. Although none of them can represent all network features, the most common way to describe the network is by measuring various features such as degree heterogeneity, clustering, and transitivity [24, 54]. Degree heterogeneity refers to the variation in the number of connections each node has [55]. Clustering measures the tendency of nodes to form clusters or tightly connected groups [45]. Transitivity quantifies the likelihood that if the connection of nodes can be transitive to other nodes [56]. Together, these features offer a comprehensive understanding of the network's topology. There are numerous network metrics that are worth exploring, including classical metrics such as centrality, betweenness, closeness, density, and clustering coefficient [5].

Based on the aforementioned degree heterogeneity, clustering, and transitivity, in this research, two indicators: DD, CC were utilized to measure the preservation of the reduced network compared to the original one. Basically, DD is a feature of a graph that shows how many edges a node is connected to [24]; CC is for measures how much nodes in a graph tend to cluster together [25]

In order to examine the preservation or differences in network structures. Goldstein proposed a different approach to measure the different between networks by using the probability distribution method to describe the network characteristics [26]. Goldstein then introduced the KS-statistic method to calculate the probability distribution gap statistically. By converting the two indicators: DD and CC into the form of probability distribution. In Ghavipour and Meybodi’s study, KS-statistic were applied to compute the distance between the true distribution of the original graph and the approximate distribution of the sampled subgraph [57]. By following Ghavipour and Meybodi’s work, this research used the KS-statistic to describe and compare the

original and reduced network based on the indicators: DD and CC. This subsection provides the detailed information of DD and CC including their definitions and calculation methods. Additionally, the calculation of KS-statistic was addressed accordingly in relation to these network measures.

Degree Distribution

DD [24] are used to measure a whole graph, which represent how many nodes have each degree. The network degree is the number of connections of a node to other nodes in the network. For example, in a social network if a person has 100 friends, then the person node has a degree of 100. DD can be represented as probability $\boldsymbol{P_{deg}(k)}$ under a certain network degree k Eq (1):

$$\boldsymbol{P_{deg}(k)} = \frac{n_k}{n} \qquad \text{Eq (1)}$$

Where $\boldsymbol{P_{deg}(k)}$ represents the probability distribution of nodes with degree k in the network, n is the total number of nodes in the graph, and n_k is the number of nodes which have degree k from 0 to maximum degree of graph. Degree refers to the number of connections (edges) a node has in a network, and the degree distribution reveals the frequency or probability distribution of these degrees across all nodes in the network, thus providing insight into the connectivity pattern.

Clustering Coefficient Distribution

Introduced in 1998 by Watts [58], CC is to quantify the level of clustering between nodes and reflects the local configuration of triangles within a graph. A high CC can indicate a collaborative relationship between nodes in a social network. It can be calculated as Eq (2):

$$\boldsymbol{CC}(i) = \frac{t_i}{T_i} = \frac{2 * t_i}{d(i)(d(i)-1)} = \frac{2 * |\{e_i : i, i \in N_i, e_i \in E\}|}{k(k-1)} \qquad \text{Eq (2)}$$

where t_i means the number of triangles, T_i means the number of all possible triangles with all its neighbors, $d(i)$ means the degree of node *i,* e_i means the edge connected to node i, N_i is a set of neighborhoods for node *i.* The formula of clustering coefficient distribution $P_{CC}(r)$ is defined as Eq (3):

$$\boldsymbol{P_{CC}(r)} = \frac{num(W|W \subset V,\ CC(W) = r,\ 0 \leq r \leq 1)}{num(V)} \quad \text{Eq (3)}$$

Where $\boldsymbol{P_{CC}(r)}$ represents the probability distribution of clustering coefficients (CC) for a specific value of r in the network, $num(W|W \subset V,\ CC(W) = r,\ 0 \leq r \leq 1)$ is the number of node subsets W that satisfy two conditions: 1) W is a subset of the entire node set V. 2) The clustering coefficient of the set W is equal to a specific value r, where $0 \leq r \leq 1$.

KS-Distance

KS-statistic is widely used to measure the difference between two cumulative distribution functions (CDF) [26]. The KS-statistic can be calculated as the maximum vertical distance between two CDFs as **Error! Reference source not found.** :

$$\boldsymbol{KS} = \max_x |F(x) - F'(x)| \quad \text{Eq (4)}$$

where x denotes the range of the random variable, F and F' are two CDFs, and $0 \leq KS \leq 1$. This method was applied to measure the distance between indicators of the original graph and the reduced subgraph. This study used this method to measure the similarity between networks based on DD, and CC

3.2. Normalized Adjusted Ratio Sampling (NARS)

For the adjusted ratio sampling optimization, the method NARS was proposed to allocate a specific number of nodes to each cluster (community) in a graph or network based on the certain considerations. NARS aims to obtain a proportional distribution of nodes while taking into account the overall graph size, the distribution of nodes among clusters, and a desired ratio of nodes selected for the reduced graph. The proposed

NARS method defines $\boldsymbol{NAS}(C_k)$ is the normalized adjusted number of nodes selected from each represented cluster (community) C_k. Basically, $\boldsymbol{NAS}(C_k)$ is the number of nodes should be selected in each cluster. The formula for $\boldsymbol{NAS}(C_k)$ can defined as Eq (5):

$$\boldsymbol{NAS}(\boldsymbol{C_k}) = \left[\left[\boldsymbol{N}(\boldsymbol{G})[\boldsymbol{0}] \times 10^{(Digit(N(G))-1)} \times \frac{\boldsymbol{max}\,(\boldsymbol{N}(\boldsymbol{C}))}{\boldsymbol{N}(\boldsymbol{G})}\right] \times \frac{\boldsymbol{N}(\boldsymbol{C_k}) - \boldsymbol{min}\,(\boldsymbol{N}(\boldsymbol{C}))}{\boldsymbol{max}\big(\boldsymbol{N}(\boldsymbol{C})\big) - \boldsymbol{min}\,(\boldsymbol{N}(\boldsymbol{C}))}\right] \quad \text{Eq (5)}$$

In Eq (5), $\boldsymbol{N}(\boldsymbol{G})$ means number of nodes in the original graph; $\boldsymbol{N}(\boldsymbol{G})[\boldsymbol{0}]$ indicates the first digit of $\boldsymbol{N}(\boldsymbol{G})$; $\boldsymbol{Digit}\big(\boldsymbol{N}(\boldsymbol{G})\big)$ means the number of digits in $\boldsymbol{N}(\boldsymbol{G})$. For instance, if $\boldsymbol{N}(\boldsymbol{G})$ is 800, then $\boldsymbol{N}(\boldsymbol{G})[\boldsymbol{0}]$ will be 8 and $\boldsymbol{Digit}\big(\boldsymbol{N}(\boldsymbol{G})\big)$ will be 3. $\boldsymbol{N}(\boldsymbol{C_k})$ is the number of nodes in cluster $\boldsymbol{C_k}$; $\boldsymbol{max}\big(\boldsymbol{N}(\boldsymbol{C})\big)$ is the number of nodes in the maxima size cluster within the original graph, and $\boldsymbol{min}\big(\boldsymbol{N}(\boldsymbol{C})\big)$ is the number of nodes in the minima size cluster within the original graph. This formula was proposed to include the fuzzy number of nodes in a cluster with maximum nodes, and it also used min-max normalization to assign varied numbers of nodes to each cluster. It considers the differences across the clusters while keeping the number of nodes chosen from the maximum cluster.

The normalized adjusted ratio sampling function, $\boldsymbol{NARS}\big(\boldsymbol{C_{k,\ ratio}}\big)$, is defined as Eq (6). Under the provided fraction of the total data collected, the formula assigned a distinct number of nodes to each cluster.

$$\boldsymbol{NARS}\big(\boldsymbol{C_{k,\ ratio}}\big) = \boldsymbol{NAS}(\boldsymbol{C_k}) \times \frac{\boldsymbol{N}(\boldsymbol{G})}{\sum \boldsymbol{NAS}(\boldsymbol{C})} \times \boldsymbol{ratio} \quad \text{Eq (6)}$$

In Eq (6), $\boldsymbol{ratio}$ is the proportion of nodes selected from the original graph to be in the reduced graph; $\sum \boldsymbol{NAS}(\boldsymbol{C})$ is the sum of $\boldsymbol{NAS}(\boldsymbol{C_k})$ in each cluster. Basically, $\boldsymbol{NARS}\big(\boldsymbol{C_{k,\ ratio}}\big)$ can be calculated by multiplying $\boldsymbol{NAS}(C_k)$ by the reciprocal of the proportion of $\sum \boldsymbol{NAS}(\boldsymbol{C})$ contained in $\boldsymbol{N}(\boldsymbol{G})$, and it is further scaled up or down by ratio. If $\boldsymbol{NARS}\big(\boldsymbol{C_{k,\ ratio}}\big)$ is less than 2, it will be 0 because, in practice, one single node in

an independent community is not helpful for us.

3.3. Overall Framework

The proposed data analysis framework is shown in

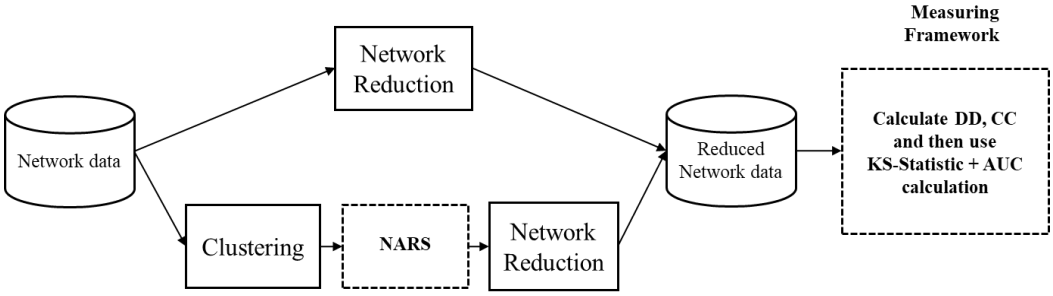

Figure 2. First, the top side of

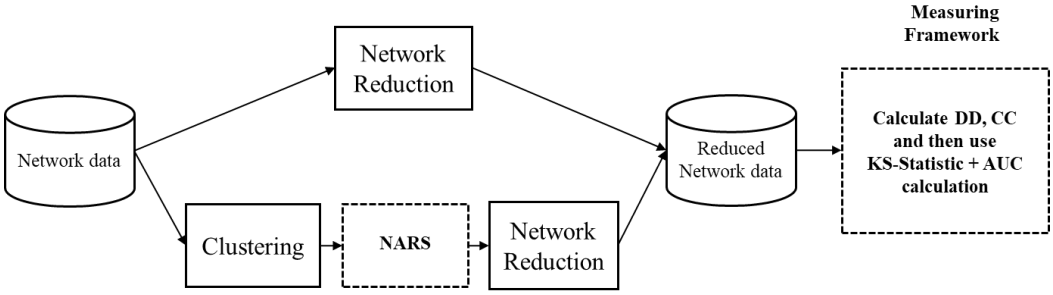

Figure 2 shows a non-clustering framework where Louvain and Leiden methods were not applied before conducting network reduction methods: RW, ISRW, and TIES. The bottom side of

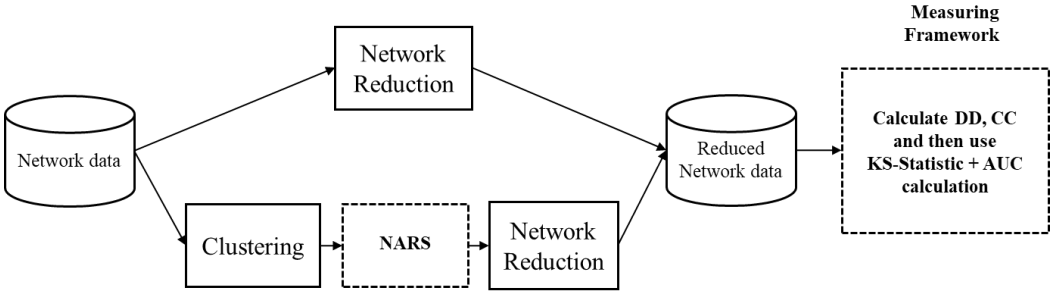

Figure 2 shows the clustering methods Louvain and Leiden were added and the proposed NARS was conducted to before applying network reduction methods. In order to compare both network graphs of non-clustering and clustering frameworks, KS-statistic distance measurements were applied to DD and CC and then do area under the

curve calculation to measure the closeness between networks. Finally, multiple experiments on different social network datasets were conducted to compare the difference between clustering and non-clustering. In the experiment, 20 different network graphs were collected from open datasets, and statistical tests were conducted to test the significance. The details of the experimental results can be found in the next chapter.

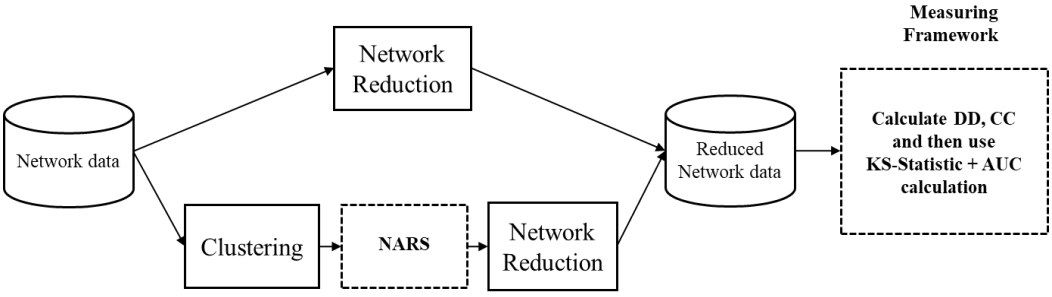

Figure 2 Proposed research framework of network reduction with clustering and comparisons

4. EXPERIMENT

This chapter presents the results by comparing different methods and provides statistical validation. The interaction between subgroups and reduction algorithms have investigated. The study uses indicators from the previous chapter and the KS-statistic distance to measure differences between graphs.

4.1. Data Description

In the experiment, 20 well-known datasets publicly were used. Their features are shown in Table 1, including nodes, edges, and average clustering coefficient (ACC), ordered by the size of nodes.

Table 1 features of 20 datasets

	Nodes	**Edges**	**ACC**
Polbooks [63]	105	441	0.4875
Jazz [64]	202	2745	0.6052
Metabolic [65]	453	2025	0.6464

VTK [66]	788	1370	0.0597
WebKB [63]	877	1388	0.1908
email-EU [67]	1005	16064	0.3994
MySQL [68]	1501	4202	0.1540
LC-multiple [69]	1536	2844	0.2925
US-airport [70]	1574	17215	0.5042
Cora [63]	2708	5278	0.2406
BTC-alpha [71]	3782	14123	0.1767
Power [58]	4941	6594	0.0801
CA-GrQC [72]	5242	14484	0.5296
BTC-otc [71]	5881	21492	0.1775
Erdos [64]	6927	11850	0.1239
Wiki [63]	7115	100762	0.1409
LastFM [63]	7624	27806	0.2194
CA-HepTh [72]	9877	25973	0.4714
PGP [64]	10680	24316	0.2659
FB [73]	22470	170823	0.3597

4.2. Performance Metrics for Network Simplification

In this study, we employed the KS-distance as a key measure to evaluate the differences between various network structures. The KS distance is a non-parametric statistical metric widely used to quantify the dissimilarity between two probability distributions. In the context of network analysis, this distance metric provides a comprehensive way to compare the DD and CC distributions of different networks. The KS distance can effectively capture the discrepancies in these distributions, allowing us to know the differences between network configurations.

We introduced the Area Under the Curve (AUC) calculation to calculate the area under the curve. The KS-distance and the AUC both fall within the range of 0 to 1 the outcomes revealed that achieving a lower score corresponds to better performance. Since the KS-distance of compared DD and CC between the initial and simplified networks the smaller value signifies enhanced results, as it implies that the simplification process has yielded a more balanced and uniform distribution across

these metrics. This, in turn, contributes to a network structure that is easier to interpret and comprehend. Consequently, when assessing AUC, a smaller value is more favorable outcome.

In summary, the decision to use KS-distance for network comparison and subsequently AUC for performance evaluation was driven by the need to capture the intricacies of network structural differences and the effectiveness of analysis methods. The combination of these metrics allows us to evaluate the impact of network reduction and clustering techniques on network analysis outcomes.

4.3. Comparison Between Different Clustering Algorithms

In order to compare the performance of Louvain's and Leiden's clustering methods, a particular data "webKB" dataset was used because “webKB” is most similar to the datasets that have been used in the literature which proposed Leiden [50]. In order to further quantify the extent of the difference, AUC calculation was used to calculate area under the curve the results are shown in Table 2 and Table 3 respectively.

Table 2 AUC with different clustering algorithm in DD

Degree Distribution		
Reduction Method	**Louvain**	**Leiden**
RW	0.06306	**0.06214**
ISRW	0.06366	**0.06312**
TIES	0.07325	**0.07206**

Table 3 AUC with different clustering algorithm in clustering CC

Clustering Coefficient Distribution		
Reduction Method	**Louvain**	**Leiden**
RW	0.08796	**0.08651**
ISRW	**0.08605**	0.08694
TIES	0.07152	**0.07131**

These tables reveal that the discrepancies are notably slight, measuring less than 0.01, across different reduction algorithms. However, it becomes evident that the

Leiden's method outperforms the Louvain's in terms of KS-distance in DD and CC and computational time according to Traag's research. Based on these findings, Leiden's method was selected as the optimal choice for subsequent research endeavors.

In addition, it's worth noting that the TIES method demands several times the computational time of the RW-based method result shown as Table 4 Time cost in different proportion for each network reduction. Given this substantial time difference, this study has opted not to delve deeper into a comprehensive comparison between these methods. Instead, the forthcoming experiment will concentrate on evaluating the RW, and ISRW methods, with a specific emphasis on the decision of whether to implement clustering or not.

Table 4 Time cost in different proportion for each network reduction

Time (Sec)	90%	80%	70%	60%	50%	40%	30%	20%	10%
RW	0.251	0.22	0.22	0.22	0.21	0.19	0.19	0.19	0.17
ISRW	0.27	0.245	0.23	0.23	0.21	0.18	0.18	0.18	0.17
TIES	9.58	6.27	4.55	3.2	2.4	1.65	1.21	0.76	0.417

4.4. Comparison Between Different Reduction Methods

In this experiment, following the main research framework outlined in

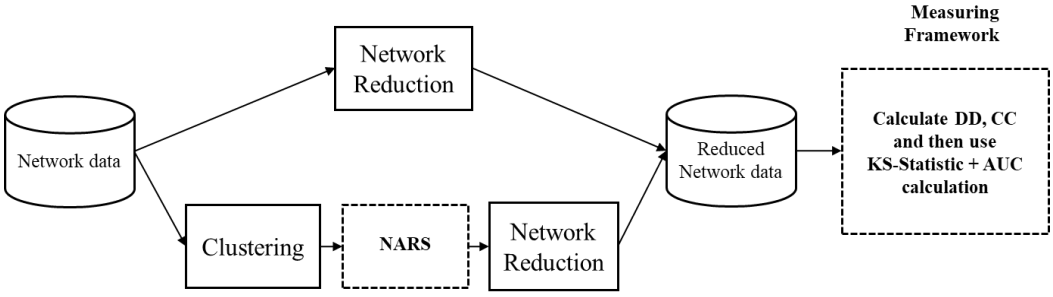

Figure 2 Proposed research framework of network reduction with clustering and comparisonsthere are two phases according to the framework. Firstly, cluster the original network by reducing the network for each cluster, and then combine all the

reduced subgraphs to calculate their indicators. Secondly, directly reduce the original network and calculate the indicators of the subgraphs without clustering.

After the reduction, the KS-distance have been calculated from the original network to the subgraphs with and without clustering, and compute the AUC from the KS-distance curve. To reduce stochastic effects and be able to use the ANOVA test, each network dataset has been calculated 50 times in each of the two frameworks, and the results will be collected and averaged. The AUC results of 20 datasets with different clustering and reduction networks are shown in Table 5 AUC and Time cost result in DD, CC for RW method with and without clustering. for RW method and Table 6 AUC and Time cost result in DD, CC for ISRW method with and without clustering. for ISRW method. The best results for a particular indicator: DD, CC, and Time in each dataset were set to bold.

Table 5 AUC and Time cost result in DD, CC for RW method with and without clustering.

	DD		CC		Time (Sec)	
	Non-cluster	Cluster	Non-cluster	Cluster	Non-cluster	Cluster
Polbooks	**0.4976**	0.5965	**0.5645**	0.6352	**0.38**	2.35
VTK	**0.0699**	0.1118	**0.0523**	0.1004	**3.07**	12.92
WebKB	**0.0438**	0.1121	**0.0303**	0.1187	**3.81**	15.82
Metabolic	**0.4094**	0.5155	0.5657	**0.5446**	**1.84**	8.77
Jazz	**0.6448**	0.6725	0.7057	**0.6879**	**1.33**	13.38
LC-multiple	**0.0806**	0.1649	**0.0621**	0.1993	**18.61**	39.58
MySQL	**0.1102**	0.2533	**0.1594**	0.2809	**8.24**	27.35
Cora	**0.1196**	0.2194	**0.1489**	0.2125	**18.03**	42.70
Power	**0.0314**	0.0937	**0.0247**	0.0480	**35.15**	63.93
Erdos	0.0840	**0.0279**	**0.0445**	0.0567	**52.21**	75.32
BTC-alpha	**0.0830**	0.1056	**0.1564**	0.1817	**46.72**	92.53
CA-GrQC	**0.0624**	0.1107	**0.1607**	0.2003	**225.90**	430.54
email-EU	**0.4455**	0.4976	0.6058	**0.5396**	**12.85**	84.65
US-airport	**0.1216**	0.2821	**0.4191**	0.4833	**32.61**	140.45
BTC-otc	**0.0855**	0.0981	**0.1503**	0.1770	**78.33**	154.18
PGP	0.0596	**0.0462**	**0.0549**	0.0723	**242.77**	706.17
CA-HepTh	**0.0800**	0.1211	0.2057	**0.1769**	378.57	**305.10**
Lastfm	**0.1215**	0.1499	**0.1881**	0.1937	**87.30**	190.04
Wiki	**0.1772**	0.1845	0.2960	**0.2638**	**351.36**	1064.62
FB	**0.2136**	0.2823	**0.3467**	0.3473	**822.90**	1457.22

Table 6 AUC and Time cost result in DD, CC for ISRW method with and without clustering.

	DD		CC		Time cost (Sec)	
	Non-cluster	Cluster	Non-cluster	Cluster	Non-cluster	Cluster
Polbooks	**0.1439**	0.2327	**0.1311**	0.2203	**0.70**	2.13
VTK	0.1199	**0.0542**	0.0971	**0.0517**	**3.83**	12.67
WebKB	**0.0602**	0.0634	**0.0519**	0.0875	**4.71**	14.55
Metabolic	**0.0651**	0.0862	**0.0551**	0.0781	**5.70**	8.20
Jazz	**0.2190**	0.2817	**0.1362**	0.1520	**6.53**	13.41
LC-multiple	0.1132	**0.1015**	0.1159	**0.1073**	**11.22**	37.90
MySQL	0.1004	**0.0820**	**0.1072**	0.1118	**13.06**	23.80
Cora	**0.0617**	0.0647	**0.0461**	0.0652	**14.81**	41.29
Power	**0.0202**	0.0388	0.0287	**0.0282**	**18.53**	39.36
Erdos	0.1342	**0.0994**	0.1279	**0.1097**	**56.27**	67.29
BTC-alpha	0.1740	**0.1542**	0.1770	**0.1733**	**59.56**	85.28
CA-GrQC	0.1296	**0.0847**	0.0969	**0.0661**	**55.63**	163.71
email-EU	0.1627	**0.1567**	0.1467	**0.1458**	81.27	**81.14**
US-airport	0.2194	**0.1884**	0.1658	**0.1624**	187.27	**136.11**
BTC-otc	0.1737	**0.1412**	0.1727	**0.1538**	152.85	**142.75**
PGP	0.1071	**0.0868**	0.1251	**0.1110**	**93.51**	160.45
CA-HepTh	0.1242	**0.0882**	0.0877	**0.0751**	**84.02**	209.43
Lastfm	0.1310	**0.1263**	**0.1522**	0.1684	**87.98**	157.19
Wiki	0.2594	**0.2512**	0.2471	**0.2449**	1178.94	**1005.62**
FB	0.1440	**0.1088**	0.1138	**0.1082**	**839.25**	1173.55

The findings from our study demonstrate an interesting aspect of RW method. It reveals that even without involving clustering method, RW can achieve superior outcomes in terms of AUC for DD and CC. This indicates that RW has capability to deliver enhanced results in these aspects without the additional cost of clustering.

In contrast, involving the ISRW method has provided us with significant insight. Specifically, we found that incorporating clustering method, particularly when dealing with larger social network datasets, which have higher edge counts, yielded notably improved outcomes. This observation gains further clarity when considering the AUC in conjunction with the KS-statistic for DD and CC analyses. The results strongly suggest that involving clustering enhances the performance of the ISRW method, particularly in scenarios involving complex and expansive social network datasets.

In short, the study has provided valuable insights into the efficacy of ISRW and

RW. When implementing ISRW, our findings strongly suggest the integration of clustering to yield improved results. This is especially true when dealing with larger-scale network datasets, where clustering not only enhances performance but also potentially contributes to better computational efficiency in some case. Conversely, the RW method results indicate that clustering might not be necessary to achieve enhanced outcomes. Even without clustering method, the RW method can deliver superior results.

4.5. Comparison between RW method and ISRW method

In this subsection, the result of the experiment comparing RW method with ISRW method is shown across all conditions, encompassing scenarios both with and without clustering. After gathering all results to Table 7, the datasets have been rearranged by the amount of edge in datasets which is the main concern in social network datasets. The columns for DD, CC, and time cost have been separated as well as clustering or non-clustering.

As can be seen in Table 7, for DD, RW method provides more “best” results (11 out of 20) than ISRW. However, for CC, it seems ISRW has more “best” results (13 out of 20). In fact, based on ANOVA statistical analysis, there is no significant difference between RW and ISRW. When considering the computational time, obviously, RW has faster performance than ISRW (14 out of 20). Particularly, when the network is larger, RW shows more efficient performance, except for PGP and CA-HepTh datasets.

From the perspective of network reduction algorithms, after applying various reduction methods, the results show that there are no differences in reduction performance when comparing RW with ISRW. For the dataset tested applying the RW method is slightly better than the ISRW method with clustering in both AUC and computational time especially when the main consideration is DD. The time cost changes drastically when combining clustering with the RW method. Final results show

no significant improvement between the community detection method and the network reduction algorithm for the RW method.

However, combining clustering with the ISRW method can yield better results when the dataset is larger. In some cases, combining clustering with ISRW can even reduce the computational time especially when CC distribution is the main concern. If the dataset is more complex and is considered a large social media dataset, using ISRW with clustering may benefit in the area of the ability to capture localized structures and relationships within networks. The combination can lead to more accurate, detailed, and representative insights compared to the more generalized nature of traditional RW.

Table 7 AUC and Time cost result in DD, CC for RW and ISRW method with and without clustering.

Dataset	DD				CC				Time cost			
	RW		ISRW		RW		ISRW		RW		ISRW	
	Non-cluster	Cluster	Non-cluster	Cluster	Non-cluster	Cluster	Non-cluster	Cluster	Non-cluster	Cluster	Non-cluster	Cluster
Polbooks	0.4976	0.5965	**0.1439**	0.2327	0.5645	0.6352	**0.1311**	0.2203	**0.38**	2.35	0.70	2.13
VTK	0.0699	0.1118	0.1199	**0.0542**	0.0523	0.1004	0.0971	**0.0517**	**3.07**	12.92	3.83	12.67
WebKB	**0.0438**	0.1121	0.0602	0.0634	**0.0303**	0.1187	0.0519	0.0875	**3.81**	15.82	4.71	14.55
Metabolic	0.4094	0.5155	**0.0651**	0.0862	0.5657	0.5446	**0.0551**	0.0781	**1.84**	8.77	5.70	8.20
Jazz	0.6448	0.6725	**0.2190**	0.2817	0.7057	0.6879	**0.1362**	0.1520	**1.33**	13.38	6.53	13.41
LC-multiple	**0.0806**	0.1649	0.1132	0.1015	**0.0621**	0.1993	0.1159	0.1073	18.61	39.58	**11.22**	37.90
MySQL	0.1102	0.2533	0.1004	**0.0820**	0.1594	0.2809	**0.1072**	0.1118	**8.24**	27.35	13.06	23.80
Cora	0.1196	0.2194	**0.0617**	0.0647	0.1489	0.2125	**0.0461**	0.0652	18.03	42.70	**14.81**	41.29
Power	0.0314	0.0937	**0.0202**	0.0388	**0.0247**	0.0480	0.0287	0.0282	35.15	63.93	**18.53**	39.36
Erdos	0.0840	**0.0279**	0.1342	0.0994	**0.0445**	0.0567	0.1279	0.1097	**52.21**	75.32	56.27	67.29
BTC-alpha	**0.0830**	0.1056	0.1740	0.1542	**0.1564**	0.1817	0.1770	0.1733	**46.72**	92.53	59.56	85.28
CA-GrQC	**0.0624**	0.1107	0.1296	0.0847	0.1607	0.2003	0.0969	**0.0661**	225.90	430.54	**55.63**	163.71
email-EU	0.4455	0.4976	0.1627	**0.1567**	0.6058	0.5396	0.1467	**0.1458**	**12.85**	84.65	81.27	81.14
US-airport	**0.1216**	0.2821	0.2194	0.1884	0.4191	0.4833	0.1658	**0.1624**	**32.61**	140.45	187.27	136.11
BTC-otc	**0.0855**	0.0981	0.1737	0.1412	**0.1503**	0.1770	0.1727	0.1538	**78.33**	154.18	152.85	142.75
PGP	0.0596	**0.0462**	0.1071	0.0868	**0.0549**	0.0723	0.1251	0.1110	242.77	706.17	**93.51**	160.45
CA-HepTh	**0.0800**	0.1211	0.1242	0.0882	0.2057	0.1769	0.0877	**0.0751**	378.57	305.10	**84.02**	209.43
Lastfm	**0.1215**	0.1499	0.1310	0.1263	0.1881	0.1937	**0.1522**	0.1684	**87.30**	190.04	87.98	157.19
Wiki	**0.1772**	0.1845	0.2594	0.2512	0.2960	0.2638	0.2471	**0.2449**	**351.36**	1064.62	1178.94	1005.62
FB	0.2136	0.2823	0.1440	**0.1088**	0.3467	0.3473	0.1138	**0.1082**	**822.90**	1457.22	839.25	1173.55

5. CONCLUSION

This study proposed an analytic framework to compare the social network before and after network reduction by investigating whether the community detection (clustering) can enhance the network reduction. In this work, for handling the clustering method, NARS method was invented to ensure that each group can obtain a fair share based on the size of the sub-network. DD, CC, and KS-distance were applied to measure the difference between the networks by using AUC calculation. The proposed framework is able to benchmarks different network reduction methods in terms of structure, clustering, and transitivity using multiple sets of data and measurements. The experimental results based on the evaluation on 20 datasets show that applying clustering before network reduction is not necessarily better than non-clustering. If the dataset is more complex and is considered large social media dataset, using ISRW with clustering may benefit on CC indicator. In general, RW can obtain a faster computational time when data is larger.

Future work could focus on two areas that are not yet unified. Firstly, regarding the measurement of network indicators and similarity, there is currently no authoritative and representative research that can broadly cover all the characteristics of the network. As the indicators that measure the network change, the results of this research may also change, which is currently unpredictable. Secondly, since 2020, the application and research of deep learning neural networks in the field of graphs have developed significantly. Classic heuristic algorithms such as modularity or other meta-heuristic methods may be surpassed by the application of deep learning with the improvement of computing capability. Therefore, the use of deep learning methods in the analysis

and comparison between community detection and network reduction is another issue worthy of further study.

REFERENCES

[1] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, "Network analysis in the social sciences," *Science,* vol. 323, no. 5916, pp. 892-895, 13 Feb 2009.

[2] O. Serrat, "Social network analysis," in *Knowledge solutions*: Springer, 2017, pp. 39-43.

[3] S. Wasserman and K. Faust, "Social network analysis: Methods and applications," 1994.

[4] D. Torgerson, "Industrialization and assessment: social impact assessment as a social phenomenon," 1980.

[5] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. June 2012: Cambridge University Press, 1994.

[6] D. Watts, "Networks, dynamics, and the small-world phenomenon," *American Journal of sociology,* vol. 105, no. 2, pp. 493-527, September 1999.

[7] D. R. White, J. Owen-Smith, J. Moody, and W. W. Powell, "Networks, fields and organizations: micro-dynamics, scale and cohesive embeddings," *Computational mathematical organization theory,* vol. 10, no. 1, pp. 95-117, 2004.

[8] E. L. Kick, L. A. McKinney, S. McDonald, and A. Jorgenson, "A multiple-network analysis of the world system of nations, 1995-1999," *Sage handbook of social network analysis,* pp. 311-327, 2011.

[9] A. Quan-Haase and B. Wellman, *Computer-mediated community in a high-tech organization* (The firm as a collaborative community: reconstructing trust in the knowledge economy). 2006, pp. 281-333.

[10] J. A. Barnes and F. Harary, "Graph theory in network analysis," *Social networks,* vol. 5, no. 2, pp. 235-244, June 1983.

[11] J. L. Moreno, *Who shall survive? Foundations of sociometry, group psychotherapy and socio-drama*. 1953.

[12] F. Harary and R. Z. Norman, *Graph theory as a mathematical model in social science* (no. 2). University of Michigan, Institute for Social Research Ann Arbor, 1953.

[13] F. Zhou, S. Malher, and H. Toivonen, "Network simplification with minimal loss of connectivity," in *2010 IEEE international conference on data mining*, 20 January 2010: IEEE, pp. 659-668.

[14] H. Oh, "Aggregation of buses for a network reduction," *IEEE Transactions on Power Systems,* vol. 27, no. 2, pp. 705-712, 09 January 2012.

[15] D. Zhang, J. Yin, X. Zhu, and C. Zhang, "Network representation learning: A survey," *IEEE transactions on Big Data,* vol. 6, no. 1, pp. 3-28, 1 March 2018.

[16] S. Arrami, W. Oueslati, and J. Akaichi, "Detection of opinion leaders in social networks: a survey," in *International conference on intelligent interactive multimedia systems and services*, 2018: Springer, pp. 362-370.

[17] Y. Liu, T. Safavi, A. Dighe, and D. Koutra, "Graph summarization methods and applications: A survey," *ACM computing surveys,* vol. 51, no. 3, pp. 1-34, 22 June 2018.

[18] C. Bhaumik, A. K. Agrawal, and P. Sinha, "Using social network graphs for search space reduction in internet of things," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 05 Septemberr 2012, pp. 602-603.

[19] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science,* vol. 286, no. 5439, pp. 509-512, 15 October 1999.

[20] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," *ACM SIGCOMM computer communication review,* vol. 29, no. 4, pp. 251-262, 30 August 1999.

[21] S. Milgram, "The small world problem," *Psychology today,* vol. 2, no. 1, pp. 60-67, 1 May 1967.

[22] R. Albert, H. Jeong, and A.-L. Barabási, "Diameter of the world-wide web," *Nature,* vol. 401, no. 6749, pp. 130-131, 09 September 1999.

[23] J. Leskovec and E. Horvitz, "Planetary-scale views on a large instant-messaging network," in *Proceedings of the 17th international conference on World Wide Web*, 21 April 2008, pp. 915-924.

[24] P. N. Krivitsky, M. S. Handcock, A. E. Raftery, and P. D. Hoff, "Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models," *Social networks,* vol. 31, no. 3, pp. 204-213, 26 May 2009.

[25] S. N. Soffer and A. Vazquez, "Network clustering coefficient without degree-correlation biases," *Physical Review E,* vol. 71, no. 5, p. 057101, 2005.

[26] M. L. Goldstein, S. A. Morris, and G. G. Yen, "Problems with fitting to the power-law distribution," *The European Physical Journal B-Condensed Matter Complex Systems,* vol. 41, no. 2, pp. 255-258, 18 June 2004.

[27] G. A. Pagani and M. Aiello, "The power grid as a complex network: a survey," *Physica A: Statistical Mechanics its Applications,* vol. 392, no. 11, pp. 2688-2700, 10 January 2013.

[28] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature,* vol. 407, no. 6804, pp. 651-654, 05 October 2000.

[29] J. P. Doye, "Network topology of a potential energy landscape: A static scale-free network," *Physical review letters,* vol. 88, no. 23, p. 238701, 23 January

2002.

[30] J. Travers and S. Milgram, "An experimental study of the small world problem," in *Social networks*: Elsevier, 1977, pp. 179-197.

[31] W. Aiello, F. Chung, and L. Lu, "A random graph model for massive graphs," in *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, 01 May 2000, pp. 171-180.

[32] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani, "Predictability and epidemic pathways in global outbreaks of infectious diseases: the SARS case study," *BMC medicine,* vol. 5, no. 1, pp. 1-13, 21 November 2007.

[33] R. Guimera and L. A. N. Amaral, "Modeling the world-wide airport network," *The European Physical Journal B,* vol. 38, no. 2, pp. 381-385, 01 March 2004.

[34] M. Boss, H. Elsinger, M. Summer, and S. Thurner 4, "Network topology of the interbank market," *Quantitative finance,* vol. 4, no. 6, pp. 677-684, 18 Aug 2004.

[35] S. M. Ashraf, B. Rathore, and S. Chakrabarti, "Performance analysis of static network reduction methods commonly used in power systems," in *2014 Eighteenth National Power Systems Conference (NPSC)*, 2014: IEEE, pp. 1-6.

[36] N. Martin, P. Frasca, and C. Canudas-de-Wit, "Large-scale network reduction towards scale-free structure," *IEEE Transactions on Network Science Engineering,* vol. 6, no. 4, pp. 711-723, 26 September 2018.

[37] S. Tsugawa and H. Ohsaki, "Benefits of bias in crawl-based network sampling for identifying key node set," *IEEE Access,* vol. 8, pp. 75370-75380, 20 April 2020.

[38] J. Kleinberg, "The small-world phenomenon: An algorithmic perspective," in *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, 2000, pp. 163-170.

[39] N. Ahmed, J. Neville, and R. R. Kompella, "Network sampling via edge-based node selection with graph induction," *Department of Computer Science Technical Reports,* January 2011.

[40] M. Plantié and M. Crampes, "Survey on social community detection," in *Social media retrieval*: Springer, 2013, pp. 65-85.

[41] B. Krishnamurthy and J. Wang, "On network-aware clustering of web clients," in *Proceedings of the conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, 28 August 2000, pp. 97-110.

[42] K. Reddy, M. Kitsuregawa, P. Sreekanth, and S. Rao, "In DNIS'02: Proceedings of the Second International Workshop on Databases in Networked Information Systems," ed. London, UK: Springer-Verlag, 2002.

[43] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly, "Complex networks as a unified framework for descriptive analysis and predictive modeling in climate

science," *Statistical Analysis and Data Mining: The ASA Data Science Journal,* vol. 4, no. 5, pp. 497-511, 16 December 2011.

[44] R. Agrawal and H. Jagadish, "Algorithms for searching massive graphs," *IEEE Transactions on Knowledge and Data Engineering,* vol. 6, no. 2, pp. 225-238, 1 April 1994.

[45] F. D. Malliaros and M. Vazirgiannis, "Clustering and community detection in directed networks: A survey," *Physics reports,* vol. 533, no. 4, pp. 95-142, 30 December 2013.

[46] S. Fortunato, "Community detection in graphs," *Physics reports,* vol. 486, no. 3-5, pp. 75-174, 17 November 2010.

[47] M. Coscia, F. Giannotti, and D. Pedreschi, "A classification for community discovery methods in complex networks," *Statistical Analysis and Data Mining: The ASA Data Science Journal,* vol. 4, no. 5, pp. 512-546, 09 September 2011.

[48] M. Crampes and M. Plantié, "A unified community detection, visualization and analysis method," *Advances in complex systems,* vol. 17, no. 01, p. 1450001, 12 March 2014.

[49] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory experiment,* vol. 2008, no. 10, p. P10008, 9 October 2008.

[50] V. A. Traag, L. Waltman, and N. J. Van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," *Scientific reports,* vol. 9, no. 1, pp. 1-12, 26 March 2019.

[51] M. M. Deza and E. Deza, "Voronoi diagram distances," in *Encyclopedia of Distances*: Springer, 2013, pp. 339-347.

[52] V. Dubois and C. Bothorel, "Transitive reduction for social network analysis and visualization," in *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, 17 October 2005: IEEE, pp. 128-131.

[53] Y. Bai, H. Ding, S. Bian, T. Chen, Y. Sun, and W. Wang, "Simgnn: A neural network approach to fast graph similarity computation," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 30 January 2019, pp. 384-392.

[54] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics,* vol. 74, no. 1, p. 47, 30 January 2002.

[55] E. Estrada, "Degree heterogeneity of graphs and networks. I. Interpretation and the “heterogeneity paradox”," *Journal of Interdisciplinary Mathematics,* vol. 22, no. 4, pp. 503-529, 2019.

[56] Z. Burda, J. Jurkiewicz, and A. Krzywicki, "Network transitivity and matrix models," *Physical Review E,* vol. 69, no. 2, p. 026106, 2004.

[57] M. Ghavipour and M. R. Meybodi, "Irregular cellular learning automata-based algorithm for sampling social networks," *Engineering Applications of Artificial Intelligence,* vol. 59, pp. 244-259, 14 January 2017.

[58] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *Nature,* vol. 393, no. 6684, pp. 440-442, 04 June 1998.

[59] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 20 August 2006, pp. 631-636.

[60] F. Gehring and P. Halmos, "Graduate Texts in Mathematics," 1977.

[61] K. J. Dooley, S. D. Pathak, T. J. Kull, Z. Wu, J. Johnson, and E. Rabinovich, "Process network modularity, commonality, and greenhouse gas emissions," *Journal of Operations Management,* vol. 65, no. 2, pp. 93-113, 18 March 2019, doi: 10.1002/joom.1007.

[62] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, "Generalized louvain method for community detection in large networks," in *2011 11th international conference on intelligent systems design and applications*, 03 January 2011: IEEE, pp. 88-93.

[63] P. Chunaev, "Community detection in node-attributed social networks: a survey," *Computer Science Review,* vol. 37, p. 100286, 21 July 2020.

[64] L. Waltman and N. J. Van Eck, "A smart local moving algorithm for large-scale modularity-based community detection," *The European physical journal B,* vol. 86, no. 11, pp. 1-14, 13 November 2013.

[65] J. Duch and A. Arenas, "Community detection in complex networks using extremal optimization," *Physical review E,* vol. 72, no. 2, p. 027104, 24 August 2005.

[66] W. J. Schroeder, L. S. Avila, and W. Hoffman, "Visualizing with VTK: a tutorial," *IEEE Computer graphics and applications,* vol. 20, no. 5, pp. 20-27, 2000.

[67] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich, "Local higher-order graph clustering," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 04 August 2017, pp. 555-564.

[68] C. R. Myers, "Software systems as complex networks: Structure, function, and evolvability of software collaboration graphs," *Physical review E,* vol. 68, no. 4, p. 046116, 20 Octobor 2003.

[69] J. M. Urquiza *et al.*, "Using machine learning techniques and genomic/proteomic information from known databases for defining relevant features for PPI classification," *Computers in biology medicine,* vol. 42, no. 6, pp. 639-650, 8 May 2012.

[70] V. Colizza, R. Pastor-Satorras, and A. Vespignani, "Reaction–diffusion processes and metapopulation models in heterogeneous networks," *Nature Physics,* vol. 3, no. 4, pp. 276-282, 04 March 2007.

[71] S. Kumar, F. Spezzano, V. Subrahmanian, and C. Faloutsos, "Edge weight prediction in weighted signed networks," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 12 December 2016: IEEE, pp. 221-230.

[72] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM transactions on Knowledge Discovery from Data,* vol. 1, no. 1, pp. 2-es, 01 March 2007.

[73] A. L. Traud, P. J. Mucha, and M. A. Porter, "Social structure of facebook networks," *Physica A: Statistical Mechanics its Applications,* vol. 391, no. 16, pp. 4165-4180, 15 August 2012.