CS 328 - Introduction to Data Science
Sparsifying Graphs While Preserving Communities

Guntas Singh Saran
guntassingh.saran@iitgn.ac.in
Indian Institute of Technology Gandhinagar
Gandhinagar, Gujarat, India

Hrriday V. Ruparel
hrriday.ruparel@iitgn.ac.in
Indian Institute of Technology Gandhinagar
Gandhinagar, Gujarat, India

ABSTRACT

In this data science project, we aim to make large graphs more manageable by utilizing graph sampling techniques and keeping only a subset of the original graph while retaining key properties. This motivates us to apply community detection algorithms and analyze graph properties before and after sparsifying. Our goal is to assess performance of various existing and newly proposed sampling techniques on various datasets with community awareness.

KEYWORDS

Graph, Community Detection, Sampling, Sparsification, Property Preservation

ACM Reference Format:
Guntas Singh Saran and Hrriday V. Ruparel. 2024. CS 328 - Introduction to Data Science Sparsifying Graphs While Preserving Communities. In *Introduction to Data Science, May 01, 2024, Gandhinagar, GJ.* ACM, New York, NY, USA, 14 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

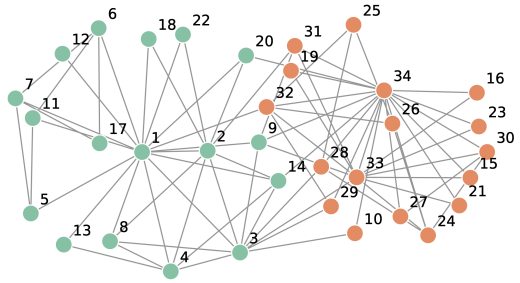


Figure 1: Communities in Zachary's Karate Club Network

Graphs can represent a wide variety of data types, including social networks, financial transactions, communication networks, and citation networks. As data size increases, it becomes challenging to analyze, store, and visualize such large network data. This is where the technique of graph sparsification becomes increasingly crucial.

Graph sparsification offers several advantages:

(1) Sparsification reduces the size of the graph, thereby saving storage space.
(2) Sparsification also motivates preservation of graph properties enabling to perform data analysis on sparsified graph.
(3) Some graph algorithms struggle with large graphs due to their high time complexity. Sparsifying the graph first can significantly reduce computation time while maintaining algorithm accuracy.
(4) When data privacy is a concern, sparsification can eliminate specific information from the graph, providing improved privacy protection.

Among multiple regimes of graph analysis algorithms, community structure detection are much sought after as they extract the close relationships of social, biological and physical networks. These algorithms have proved to be useful in multiple domains: targeted advertising, identifying influential nodes, protein-protein interactions, brain connectivity networks and recommender systems.

This project explores the performance of various existing and newly proposed graph sparsifying/sampling techniques coupled with community detection algorithms applied on some common community-aware graph datasets. We propose a unified pipeline for implementing sparsification techniques and community detection algorithms, and visualizing the performance based on defined metrics.

2 PROBLEM STATEMENT

For the purpose of graph sparsification preserving communities, we propose the following broad problem statement:

*Given a graph $G(V, E)$, where V is the vertex set and E is the edge set, we wish to **sparsify** the graph in a meaningful manner and obtain a graph $G'(V, E')$, where V is the vertex set similar to original graph G and $E' \subset E$ is the edge set of sparsified graph G'*

By **sparsifying** a graph in a meaningful manner, we intend to preserve the community structure of the original graph G by sampling a subset of edges from G so that community detection algorithms can be implemented on the sparsified graph G' with minimum loss of generality and accuracy. A pipeline for the said task has been depicted in Figure 2.

Introduction to Data Science, May 01, 2024, Gandhinagar, GJ

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

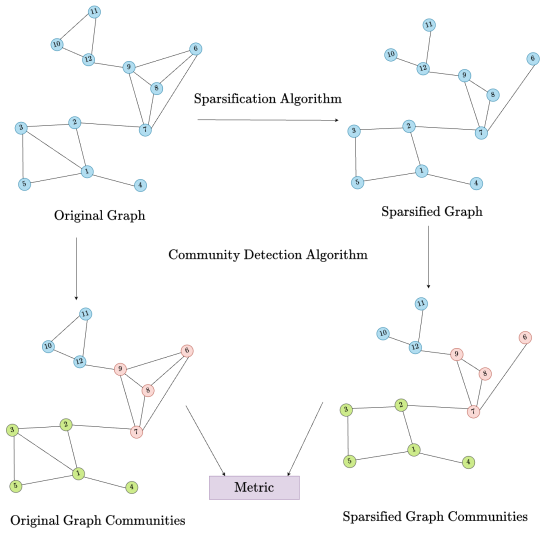


Figure 2: Evaluation Pipeline for comparing the communities produced by the Original vs. Sparsified Graph

3 SALIENT GRAPH PROPERTIES

While dealing with community detection, a multitude of pivotal properties contribute significantly to the detection and preservation of communities. These properties are fundamental for understanding the underlying structure and organization of networks, thereby aiding in the identification and characterization of cohesive communities within a given graph. Some of these salient properties include:

(1) **Edge Betweenness**: The number of shortest paths between pairs of nodes in a graph that pass through a particular edge. It quantifies the importance of an edge in connecting different parts of the network.

$$\text{Betweenness}(e) = \sum_{s,t \in V} \frac{\sigma(s,t|e)}{\sigma(s,t)} \tag{1}$$

e: Edge in graph G(V, E)
$\sigma(s,t)$: Number of shortest (s-t) paths
$\sigma(s,t|e)$: Number of shortest (s-t) paths passing through e

(2) **Jaccard Similarity** [7]: Measures similarity between two sets by comparing their intersection to their union. It is often used to measure the similarity between the sets of neighbors of two nodes.

$$\text{Sim}(i,j) = \frac{|\text{Adj}(i) \cap \text{Adj}(j)|}{|\text{Adj}(i) \cup \text{Adj}(j)|} \tag{2}$$

i, j: Nodes belonging to V
Adj(i): Adjacency list of node i

(3) **Modularity** [3] [2]: Quantifies the degree to which a network is partitioned into communities or modules. Compares the number of edges within communities to the number of edges expected in a random network with the same degree

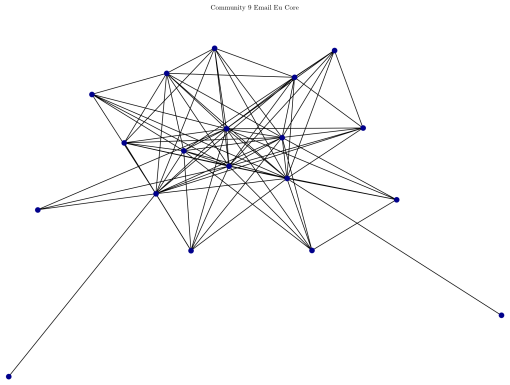


Figure 3: A Community in Email EU Core Network

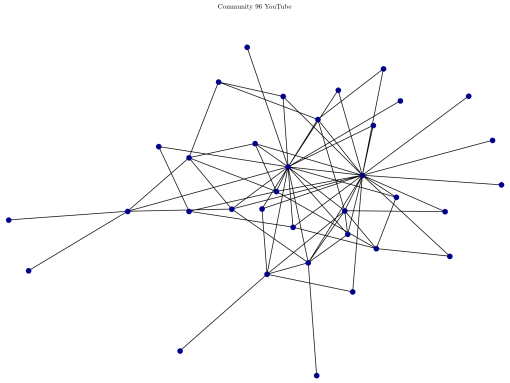


Figure 4: A Community in YouTube Social Network

Figure 5: Communities in Large Networks

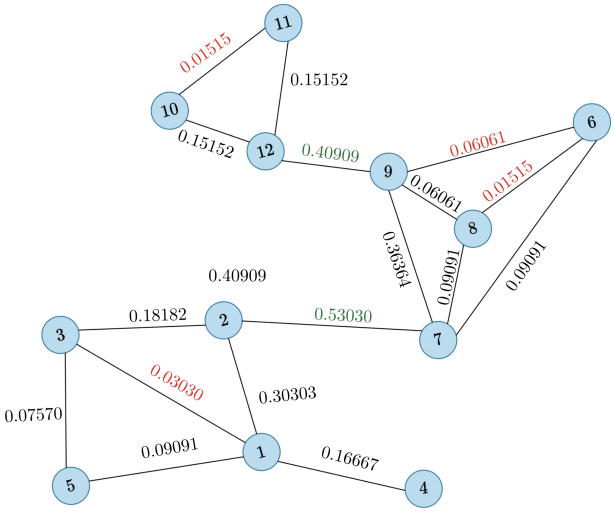


Figure 6: Edge Betweenness of All Edges in the Graph

distribution, indicating the presence of densely connected communities.

$$Q(C) = \frac{1}{2m} \sum_{C \in \mathcal{C}} \sum_{u \in C, v \in C} \left(A_{u,v} - \frac{d_u d_v}{2m} \right) \tag{3}$$

C: Set of communities
C: Community in G(V, E),
m: Number of edges in G(V, E)
d_u: Degree of node u in random rewired network G'(V, E')
$A_{u,v}$: Entry at (u,v) location of adjacency matrix of G; 1 if edge exists, else 0

4 SAMPLING TECHNIQUES

The evaluation methodology for assessing the efficacy of sparsification involves comparing the outcomes of community detection obtained from the original network with those derived from sparsified networks utilizing the same community detection algorithm.

A sparsification approach is deemed effective if the community detection outcomes from the sparsified networks closely match or even surpass those from the original network. Such sparsified networks offer benefits such as reduced storage requirements and enhanced computational efficiency for community detection algorithms, all while preserving the quality of the detected communities.

Descriptions of sampling techniques used to achieve sparsification are listed as below:

(1) **Edge Betweenness Sparsification** [Proposed]: An edge sampling algorithm that computes the edge betweenness centrality measure of each edge in G(V, E) and removes top k% of edges on the basis of centrality measure.
(2) **Random Edge Sampling** [3]: An edge sampling algorithm that uniformly retains k% of edges.
(3) **Edge Jaccard Sparsification** [7]: An algorithm that computes similarity of end points of each edge in the graph. It then sorts all the edges by their similarities and returns top k% of them.
(4) **Local Sparsification - LSpar** [7]: An algorithm that picks top d_i^e edges incident on node i (degree of node i is d_i) according to similarity (Eqn. 2).
(5) **Clustering Coefficients** [Proposed]: An algorithm that picks top k% of those edges joining nodes with maximum clustering coefficients products.

Figure 7: Facebook Socials with 50% edges retained

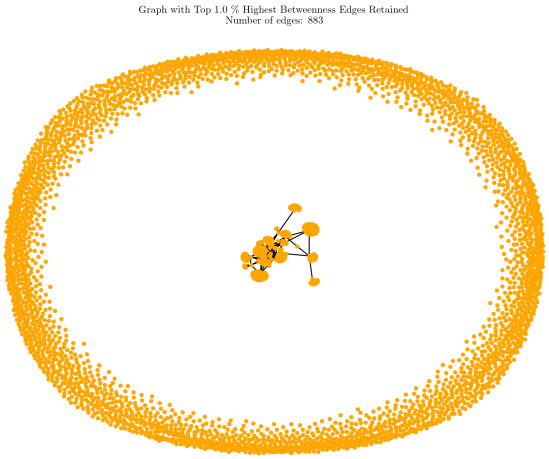


Figure 8: Facebook Socials with 1% edges retained

Figure 9: Edge Sampling on Facebook Network using High Edge Betweenness Sampling

Algorithm 1 Edge Betweenness Sparsification

```
procedure EdgeBetweennessSparsification(G, k)
    edge_betweenness ← EdgeBetweenness(G)
    edges_to_remove ← SortEdgesDec(edge_betweenness, k, G)
    H ← G.copy()
    for edge in edges_to_remove do
        H.remove_edge(edge)
    end for
    return H
end procedure
```

Algorithm 2 Random Edge Sampling

```
procedure RandomEdgeSampling(G, k)
    edges ← List of edges of G
    Shuffle edges randomly
    H ← G.copy()
    for i from 1 to ⌊(1 − k) × G.number_of_edges()⌋ do
        H.remove_edge(edges[i])
    end for
    return H
end procedure
```

Algorithm 3 Edge Jaccard Sparsification

procedure EdgeJaccardSparsification(G, k)
  $G_{\text{sparse}} \leftarrow \emptyset$
  for each edge $e = (i, j)$ in E **do**
    e.sim $\leftarrow$ Sim(i, j) according to Eqn 2
  end for
  Sort all edges in E by e.sim
  Add the top k% edges to G_{sparse}
  return G_{sparse}
end procedure

Algorithm 4 Local Sparsification Algorithm

procedure LocalSparsification($G = (V, E)$, e)
  $G_{\text{sparse}} \leftarrow \emptyset$
  for each node i in V **do**
    Let d_i be the degree of i
    Let E_i be the set of edges incident to i
    for each edge $e = (i, j)$ in E_i **do**
      e.sim $\leftarrow$ Sim(i, j) according to Eqn. 1
    end for
    Sort all edges in E_i by e.sim
    Add top d_e^i edges to G_{sparse}
  end for
  return G_{sparse}
end procedure

Algorithm 5 Clustering Coefficients Based Edge Sampling

procedure ClusteringCoeffsEdgeSampling(G, k)
  $H \leftarrow$ Empty graph
  edge_cc_prod $\leftarrow$ Empty list
  edges $\leftarrow$ Empty list
  cc $\leftarrow$ clustering_coefficients(G)
  for each edge in G.edges() **do**
    edges.append(edge)
    edge_cc_prod.append(cc[edge[0]] $\times$ cc[edge[1]])
  end for
  indices $\leftarrow$ argsort(*edge_cc_prod*)[:: -1]
  sampled_edges $\leftarrow$ edges[indices][: int(G.number_of_edges() * k)]
  H.add_edges_from(*sampled_edges*)
  H.add_nodes_from(G.nodes)
  return H
end procedure

5 METRICS FOR EVALUATION

Metrics used for evaluation of the performance of community detection algorithms before and after sparsification are as follows:

(1) **Adjusted Rand Index** [4]: The Adjusted Rand Index (ARI) measures the similarity between two data clusterings. The Adjusted Rand Index takes into account the fact that some agreement between two clusterings can occur by chance, and it adjusts the Rand Index to account for this possibility. Let a be the number of pairs of samples that are in the same cluster in both $C1$ and $C2$ and let b be the number of pairs of samples that are in different clusters in $C1$ and $C2$. We calculate the expected value E of the Rand Index for random clusterings. n_i is the number of samples in cluster i and n_j is the number of samples in cluster j.

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}} \tag{4}$$

(2) **Normalized Mutual Information** [5]: Normalized Mutual Information (NMI) is a normalization of the Mutual Information (MI) score to scale the results between 0 (no mutual information) and 1 (perfect correlation).

$$I_X^{(S)}(c; g) = \frac{I_X(c; g)}{\frac{1}{2}\left[I_X(c; c) + I_X(g; g)\right]} \tag{5}$$

where,

$$I_0(c; g) = \log \frac{n! \prod_{rs} n_{rs}^{(cg)}!}{\prod_r n_r^{(c)}! \prod_s n_s^{(g)}!} \tag{6}$$

6 NETWORKS USED

The following standard community aware network datasets were used:

(1) **DBLP** [6]: This is also a co-authorship network. Communities are determined by publication venues.
(2) **Amazon Co-Purchase Network** [6]: Edges in this network represent pairs of frequently co-purchased products. Communities represent product categories as provided by Amazon.
(3) **EU Email Core Network** [6]: This network represents the "core" of the email-EuAll network. Each community is a department in the institute such that each individual belongs to exactly one of 42 departments at the research institute.
(4) **Facebook Circles** [6]: Facebook ego-network. communities are social-circles of users.

Table 1: Network Details

Network	|V|	|E|	|C|	$\|V\|_{\text{ind}}$	$\|E\|_{\text{ind}}$
DBLP	317080	1049866	150	1420	4609
Amazon	334863	925872	300	2008	5960
EU	1005	16064	42	1005	16064
Facebook	4039	88234	NA	4039	88234

7 COMMUNITY DETECTION ALGORITHMS

Community Detection Algorithms play a crucial role in many network analysis tasks. For the purpose of this research, the detection algorithms used are listed below with short descriptions:

(1) **Louvain Algorithm** [1]: Greedy modularity-based approach for community detection in networks. It greedily optimizes

modularity by iteratively moving nodes to neighboring communities and the community structure identified is aggregated into a new network, and the process is repeated. The algorithm iterates until a maximum in modularity is reached, resulting in a partition of the network into communities.

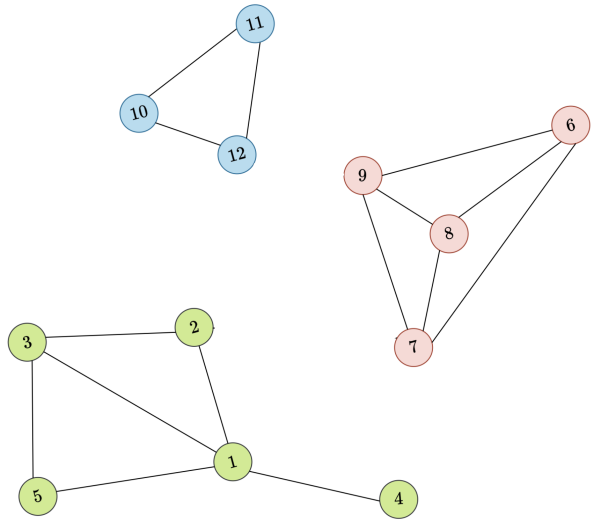


Figure 10: Partition of the graph with Max Modularity

(2) **Label Propagation Algorithm**: Method for community detection in networks. It operates by iteratively updating the labels (or community assignments) of nodes based on the majority label among their neighbors. Initially, each node is assigned a unique label, and in each iteration, nodes adopt the label that is most prevalent among their neighbors.

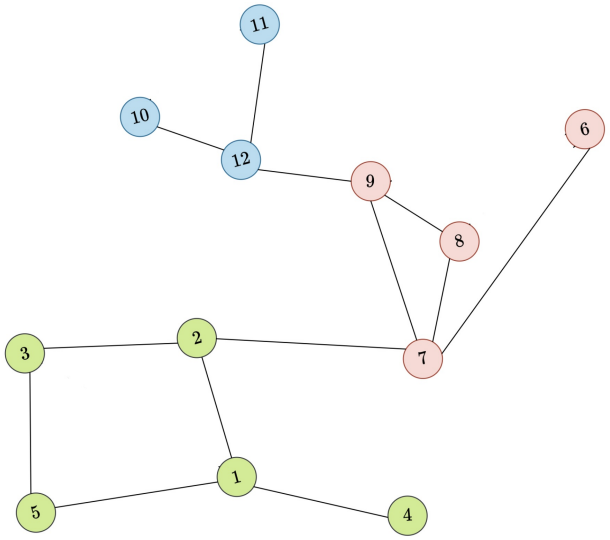
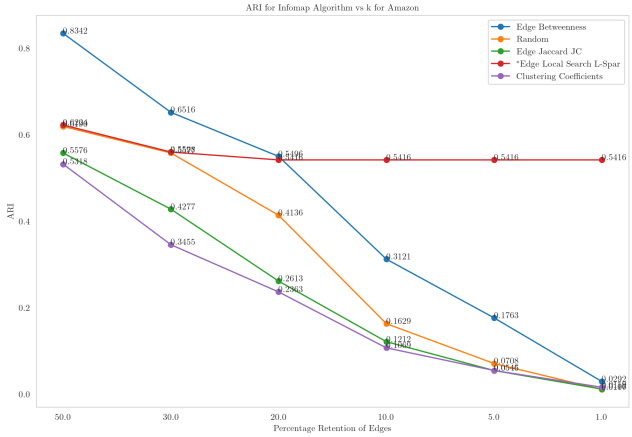


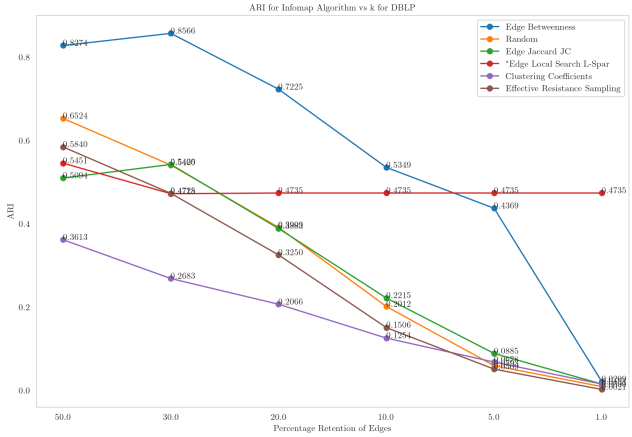
Figure 11: Communities Detected in the Sparsified Graph

(3) **InfoMap Algorithm**: Algorithm inspired by principles of information theory. Aims to find the most efficient compression of information flow in a network. It treats the network as a flow of random walks and seeks to partition it into modules that minimize the amount of information needed to describe the flow.

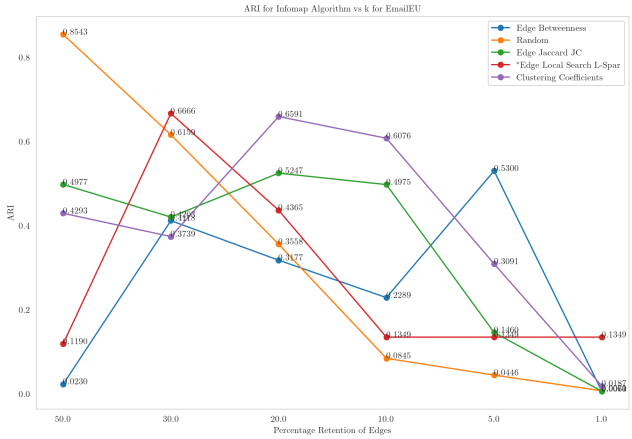
8 RESULTS



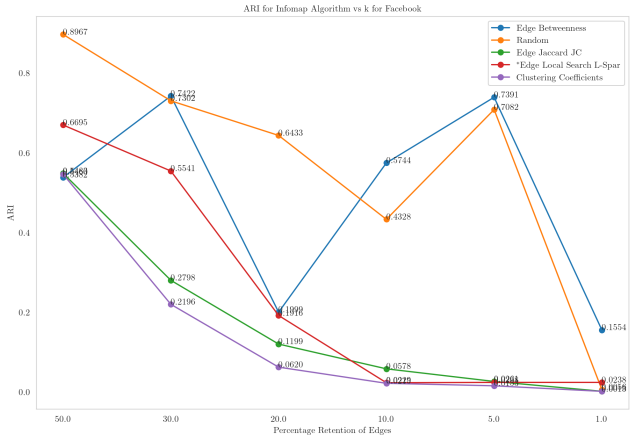
(a) Amazon Co-Purchase Network



(b) DBLP Network

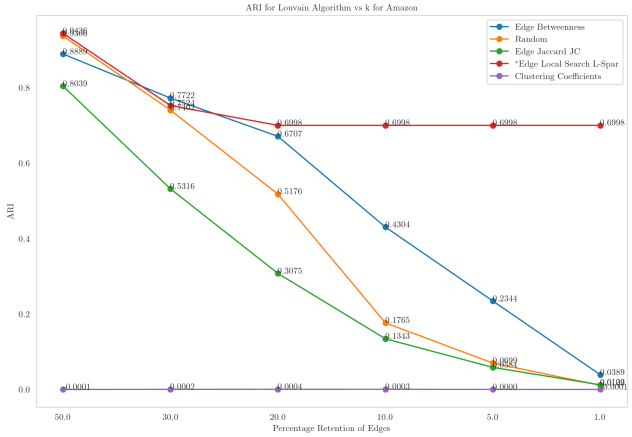


(c) Email EU Core Network

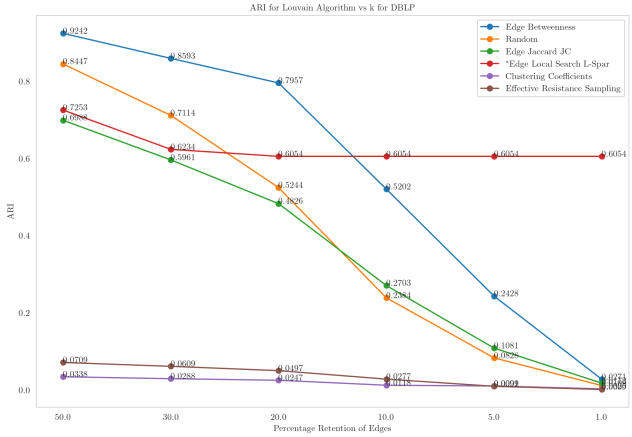


(d) Facebook Socials Network

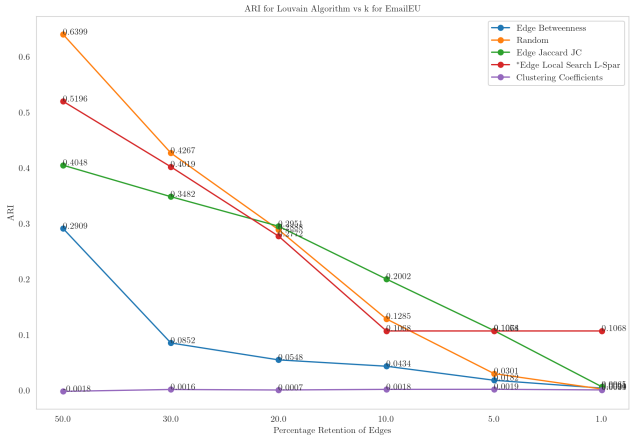
Figure 12: Adjusted Rand Index (ARI) for InfoMap Algorithm detected communities for various sparsification techniques



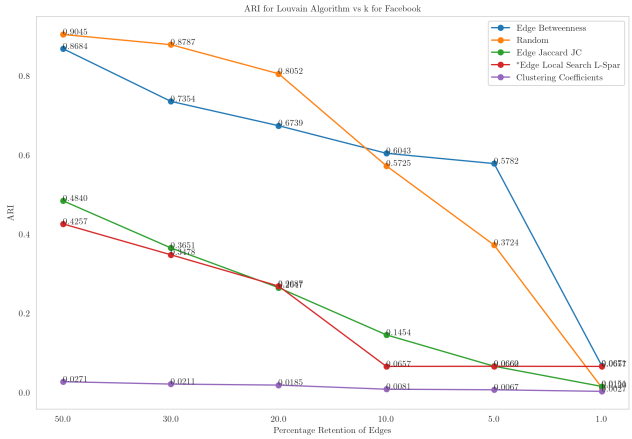
(a) Amazon Co-Purchase Network



(b) DBLP Network

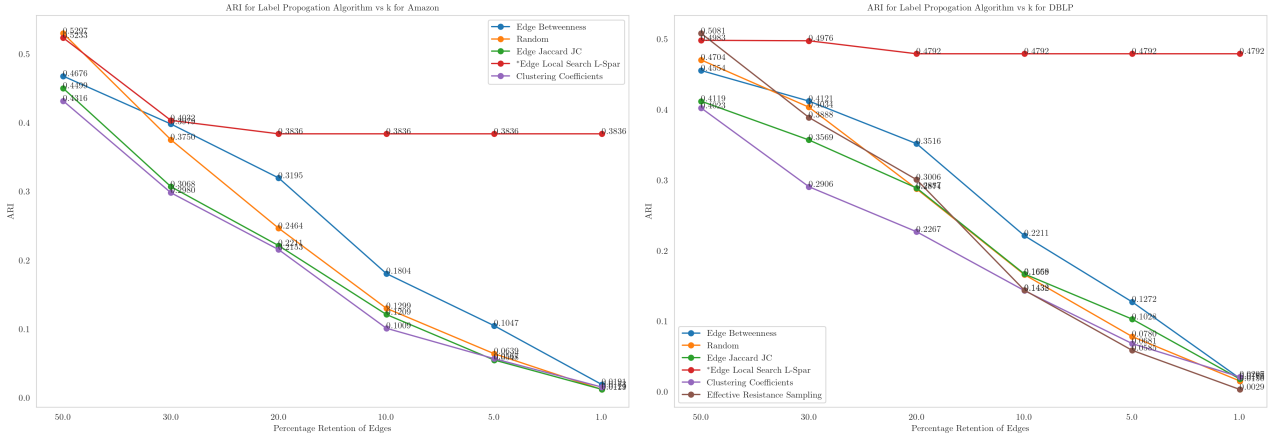


(c) Email EU Core Network



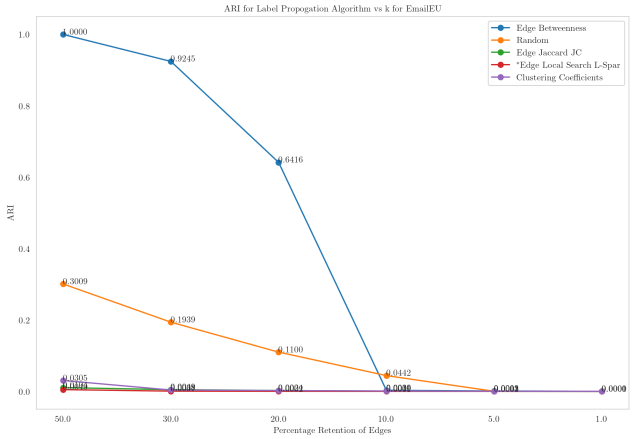
(d) Facebook Socials Network

Figure 13: Adjusted Rand Index (ARI) for Louvain Algorithm detected communities for various sparsification techniques

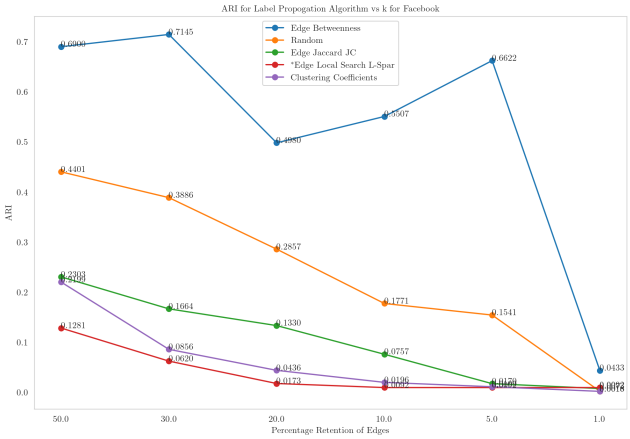


(a) Amazon Co-Purchase Network

(b) DBLP Network

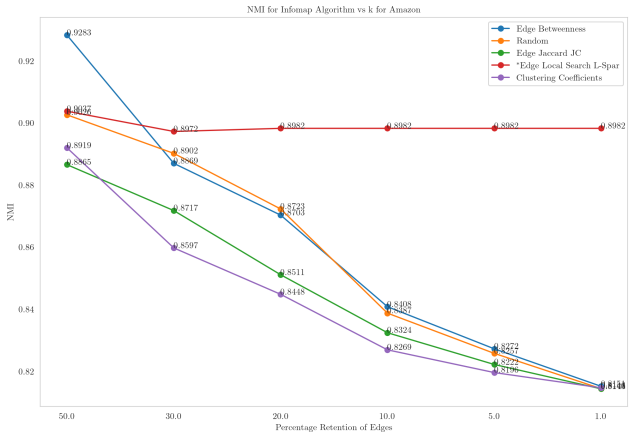


(c) Email EU Core Network

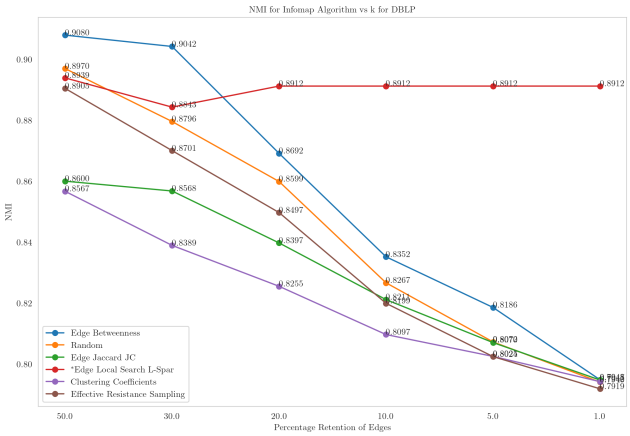


(d) Facebook Socials Network

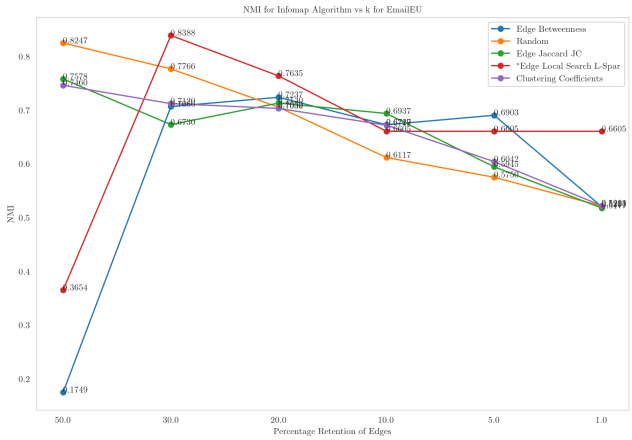
Figure 14: Adjusted Rand Index (ARI) for Label Propagation Algorithm (LPA) detected communities for various sparsification techniques



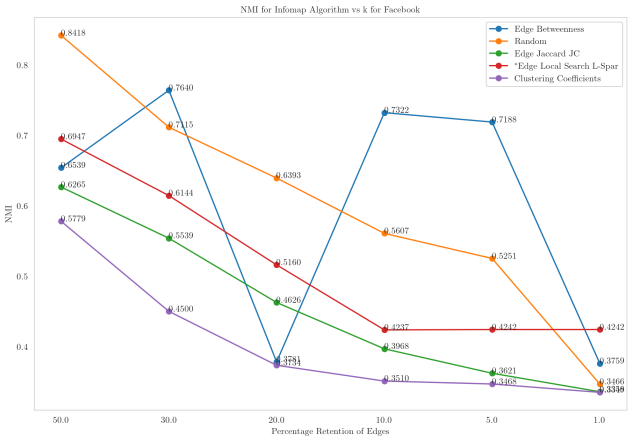
(a) Amazon Co-Purchase Network



(b) DBLP Network

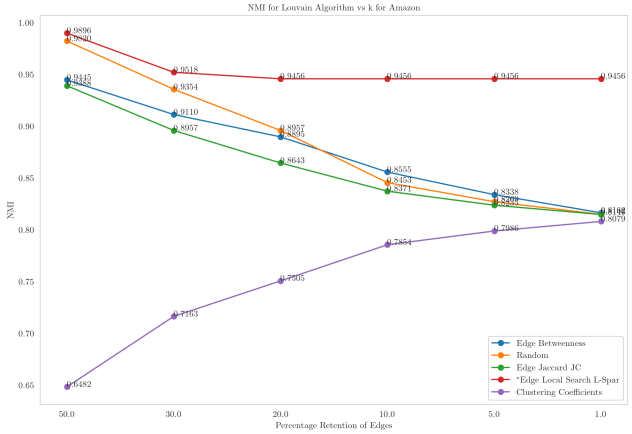


(c) Email EU Core Network

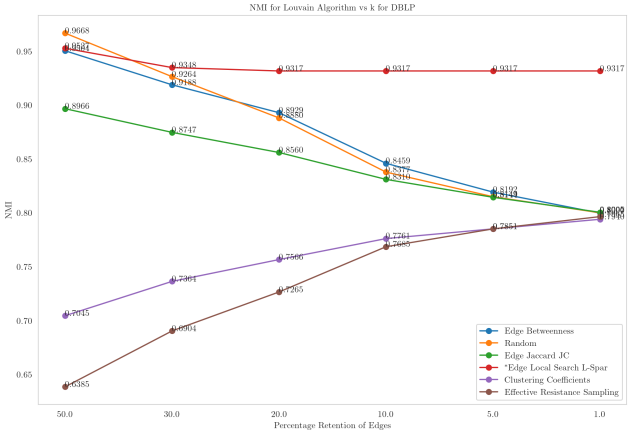


(d) Facebook Socials Network

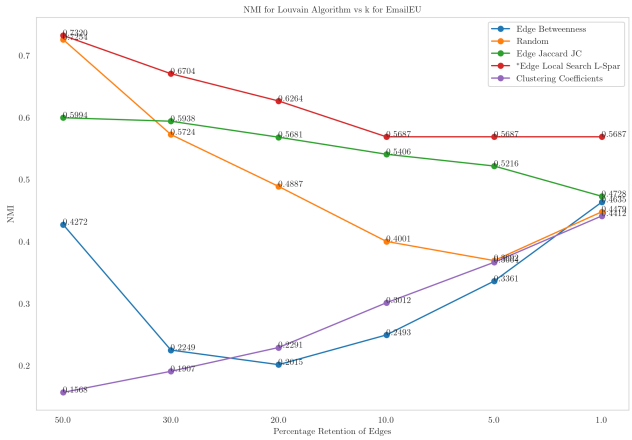
Figure 15: Normalised Mutual Information (NMI) for InfoMap Algorithm detected communities for various sparsification techniques



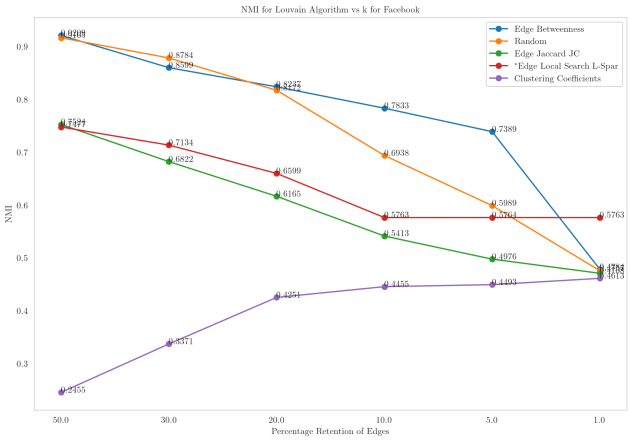
(a) Amazon Co-Purchase Network



(b) DBLP Network

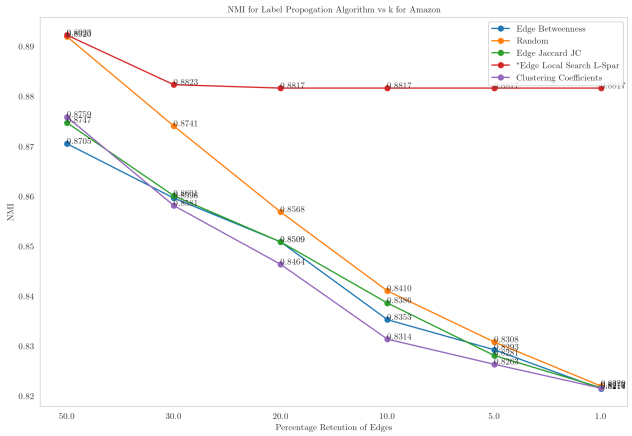


(c) Email EU Core Network

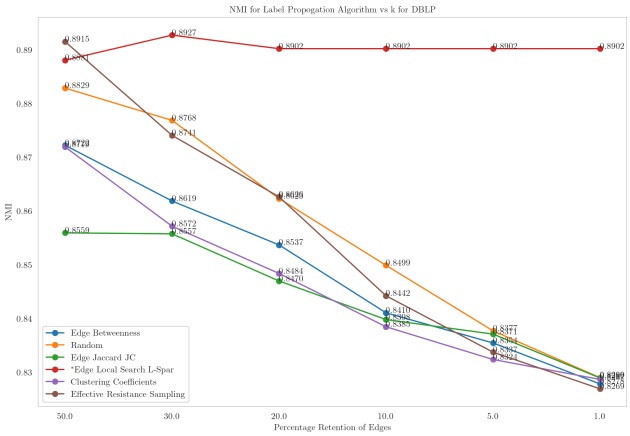


(d) Facebook Socials Network

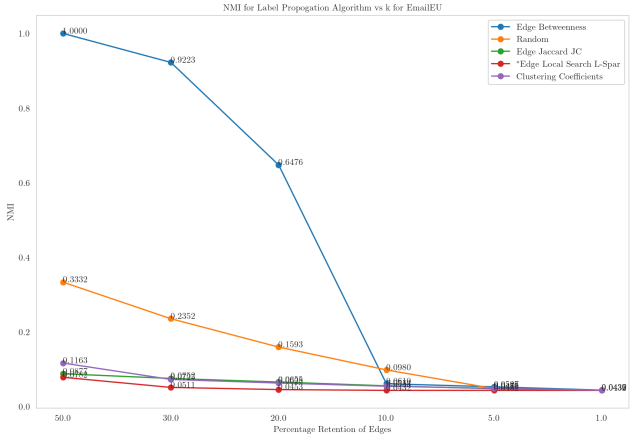
Figure 16: Normalised Mutual Information (NMI) for Louvain Algorithm detected communities for various sparsification techniques



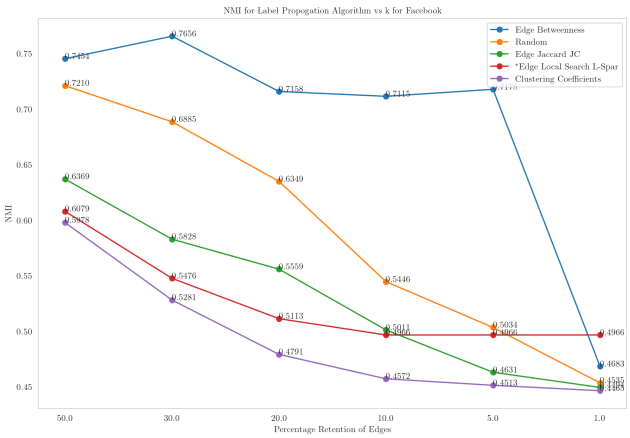
(a) Amazon Co-Purchase Network



(b) DBLP Network



(c) Email EU Core Network



(d) Facebook Socials Network

Figure 17: Normalised Mutual Information (NMI) for Label Propagation Algorithm (LPA) detected communities for various sparsification techniques

9 ACKNOWLEDGEMENTS

We extend our sincere gratitude to Prof. Anirban Dasgupta for his invaluable guidance and support throughout this project. His expertise and encouragement were instrumental in our success.

A APPENDIX: TABULAR RESULTS

REFERENCES

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (Oct. 2008), P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008

[2] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. 2008. On Modularity Clustering. *IEEE Transactions on Knowledge and Data Engineering* 20, 2 (2008), 172–188. https://doi.org/10.1109/TKDE.2007.190689

[3] Pili Hu and Wing Cheong Lau. 2013. A Survey and Taxonomy of Graph Sampling. arXiv:1308.5865 [cs.SI]

[4] Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification* 2 (1985), 193–218.

[5] Maximilian Jerdee, Alec Kirkley, and M. E. J. Newman. 2023. Normalized mutual information is a biased measure for classification and community detection. arXiv:2307.01282 [cs.SI]

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

[7] Venu Satuluri, Srinivasan Parthasarathy, and Yiye Ruan. 2011. Local graph sparsification for scalable clustering. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data* (Athens, Greece) *(SIGMOD '11)*. Association for Computing Machinery, New York, NY, USA, 721–732. https://doi.org/10.1145/1989323.1989399

Table 2: Adjusted Rand Index (ARI) for Different Sparsifying Techniques

Network	Algorithm	Technique	Metric					
			Adjusted Rand Index (ARI)					
		Degree of Sparsification (%)	50	30	20	10	5	1
Amazon	Louvain	Edge Betweenness	0.8889397150193233	0.7721709987118094	0.6707056504638057	0.4303784712626356	0.2343911173885915	0.038860193909577616
Amazon	Louvain	Random	0.936621040493242	0.7401961671250925	0.5176382573765409	0.1764715396512654	0.06989093699684017	0.010927477211447665
Amazon	Louvain	Edge Jaccard JC	0.8038758185615879	0.5316109608807337	0.30745172241406776	0.1343032101586413	0.058411970060090114	0.012173859223010002
Amazon	Louvain	Edge Local Search L-Spar	0.9435680818218039	0.7524488425336769	0.6998232039421973	0.6998232039421973	0.6998232039421973	0.6998232039421973
Amazon	Louvain	Clustering Coefficients	-0.00010088501478589729	0.00024975773038696274	0.00035687558048092395	0.0003389235963173657	1.5343638247120071e-05	-9.252003652633094e-05
Amazon	InfoMap	Edge Betweenness	0.8341674907613372	0.651608176978729	0.5496344508359453	0.3120839591095056	0.17631044353531225	0.029197746195722348
Amazon	InfoMap	Random	0.6189532047638139	0.55770320861182	0.4136277750921188	0.16288806148354895	0.07075588633443616	0.010741519213214151
Amazon	InfoMap	Edge Jaccard JC	0.5575689984702015	0.42768735100645466	0.2612716090881117	0.12123059917039097	0.05454327332674843	0.011633817597072695
Amazon	InfoMap	Edge Local Search L-Spar	0.6223772153254605	0.55975792326044	0.5415868126292722	0.5415868126292722	0.5415868126292722	0.5415868126292722
Amazon	InfoMap	Clustering Coefficients	0.5317743886901218	0.3455438727517247	0.2362711558452026	0.10691731075923337	0.05463340267029931	0.01589171885783496
Amazon	LPA	Edge Betweenness	0.4675681675154847	0.39791587980585486	0.3194608915448554	0.18040339681089895	0.10467656204988814	0.019063379801453235
Amazon	LPA	Random	0.5296820155195386	0.3750297123864478	0.24636615651780125	0.12992512063280687	0.06393133800361779	0.01189275375780645
Amazon	LPA	Edge Jaccard JC	0.44989780746277397	0.30680584384663856	0.22106629149700108	0.12092870812134095	0.05476886186609059	0.012082292342574125
Amazon	LPA	Edge Local Search L-Spar	0.5232649746619548	0.40317683307322444	0.3836372635450907	0.3836372635450907	0.3836372635450907	0.3836372635450907
Amazon	LPA	Clustering Coefficients	0.4315502271812319	0.2979689802061186	0.21525838992936122	0.10091310675650982	0.056701661152473534	0.015349219353745525
		Degree of Sparsification (%)	50	30	20	10	5	1
DBLP	Louvain	Edge Betweenness	0.9241903302919282	0.8592908804823601	0.7956618462710946	0.5202271581567174	0.24281584688560798	0.027125699993076565
DBLP	Louvain	Random	0.8446704753750868	0.7114346688027677	0.5243749028990516	0.23841414497984717	0.08278036013353306	0.010922866778860874
DBLP	Louvain	Edge Jaccard JC	0.6988080544998669	0.5960777253219467	0.48264224566938735	0.27031358870467537	0.10809846693657017	0.01720412555576489
DBLP	Louvain	Edge Local Search L-Spar	0.7253428715663303	0.6234327288234481	0.6053674761384193	0.6053674761384193	0.6053674761384193	0.6053674761384193
DBLP	Louvain	Clustering Coefficients	0.03379159748083295	0.028822863474589146	0.024731710657399292	0.011849521531440244	0.009920026649753068	0.002481414859786409
DBLP	Louvain	Effective Resistance Sampling	0.07089206462622764	0.06087390770949368	0.04969776386837865	0.027715287915034442	0.009107305687596625	0.0008636544665177169
DBLP	InfoMap	Edge Betweenness	0.8274155066619698	0.8565630893451738	0.7224930067895513	0.534907976003983	0.4368960293140232	0.020913868780493134
DBLP	InfoMap	Random	0.6524191711910338	0.5405217869503522	0.39091566593520677	0.201215127834792 12	0.05925530228525456	0.008986378767821815
DBLP	InfoMap	Edge Jaccard JC	0.5094277083704063	0.5420466590883575	0.3882551641719562	0.2215243174485267	0.08852866306228192	0.015063894200597325
DBLP	InfoMap	Edge Local Search L-Spar	0.5450565191722931	0.47183610390324693	0.4735122035514799	0.4735122035514799	0.4735122035514799	0.4735122035514799
DBLP	InfoMap	Clustering Coefficients	0.3613248704258893	0.2682574807907429	0.2065774494904706	0.12537389658284678	0.06784363774128101	0.015063894200597325
DBLP	InfoMap	Effective Resistance Sampling	0.5839556360410562	0.4725025148857017	0.3249555835857645	0.15057567728909033	0.05089138656337031	0.0021110491771470148
DBLP	LPA	Edge Betweenness	0.4554374781240136	0.4120764104433539	0.351617805470274	0.2211432848902192	0.12721027354384035	0.01870982559739486
DBLP	LPA	Random	0.4704248903917249	0.40342147604798795	0.2874440744852091	0.16592975173289454	0.0780424135573408	0.014980603756893546
DBLP	LPA	Edge Jaccard JC	0.41187298665439487	0.3569111999521394	0.28869689890914535	0.1668231569587319	0.10279199870615371	0.01787287590124668
DBLP	LPA	Edge Local Search L-Spar	0.4982791481675458	0.4976204420454734	0.4791836889671824	0.4791836889671824	0.4791836889671824	0.4791836889671824
DBLP	LPA	Clustering Coefficients	0.4022771808377241	0.2905781419562405	0.22668873775524237	0.14322982970662673	0.06809116983906653	0.02073272598139845
DBLP	LPA	Effective Resistance Sampling	0.5081104265380347	0.3888482019020112 5	0.3005501978409754	0.14380013310693063	0.05849186252646 73	0.00294807440672 3264
		Degree of Sparsification (%)	50	30	20	10	5	1
Email EU	Louvain	Edge Betweenness	0.29091401947375783	0.08522672357800397	0.05481668729236831	0.04343872123304919	0.018203804515703092	0.0040579477479807165
Email EU	Louvain	Random	0.639903780053582	0.42673710535815856	0.28881669761961876	0.12851098227022204	0.03013508555705833	0.0023666570140729514
Email EU	Louvain	Edge Jaccard JC	0.40476203793001314	0.3482243832961259	0.2950940987238714 3	0.2001703233843082	0.10743350425535311	0.006464414388151797
Email EU	Louvain	Edge Local Search L-Spar	0.519639218935995	0.40191238914780425	0.27718939004538845	0.10683989659405807	0.10683989659405807	0.10683989659405807
Email EU	Louvain	Clustering Coefficients	-0.0017705255378830673	0.0016106414245441458	0.0007139683956666705	0.0017998373638886194	0.0019325278047237012	0.0008506803594570493
Email EU	InfoMap	Edge Betweenness	0.023011948734285814	0.41178717254986624	0.31769852284838657	0.22887462145232154	0.5300274915786629	0.00602567817230517
Email EU	InfoMap	Random	0.8542599099208678	0.6159142859996335	0.3557875490769754	0.08451082447533806	0.0446357236169607	0.007394110805937746
Email EU	InfoMap	Edge Jaccard JC	0.4977058489384387	0.42027155737826644	0.5246745727977891	0.4974767124890372	0.14600269743421215	0.006257725629567975
Email EU	InfoMap	Edge Local Search L-Spar	0.11904258335261583	0.6666424799970755	0.4365095246833731	0.1349490875685066	0.1349490875685066	0.1349490875685066
Email EU	InfoMap	Clustering Coefficients	0.4292671763029602	0.3738744902051165	0.659076363308563	0.6076165125256714	0.30905157855104387	0.018653207173725546
Email EU	LPA	Edge Betweenness	1.0	0.9245201492297322	0.6415816428965775	0.003045966265386999	0.001120767519462269	3.9991513392636146e-05
Email EU	LPA	Random	0.3009085444116541	0.19387420322785892	0.10999012496376552	0.04418730140690537	0.0002174172529762512	3.1650257068040476e-05
Email EU	LPA	Edge Jaccard JC	0.010432887877614573	0.004906200421305359	0.0023856918531051227	0.0009461697185369934	0.00032372625157322507	3.751985104782929e-05
Email EU	LPA	Edge Local Search L-Spar	0.004909245984039305	0.0006700439820435201	0.00020193757156272894	0.0001041996027597066	0.0001041996027597066	0.0001041996027597066
Email EU	LPA	Clustering Coefficients	0.03054472193746821	0.0037680071176411857	0.0023549424272 19308	0.0011231470498880374	0.0005098579840945301	5.6833725432351935e-05
		Degree of Sparsification (%)	50	30	20	10	5	1
Facebook	Louvain	Edge Betweenness	0.8684112846691824	0.7354031390747675	0.6739321996754213	0.6042788561038309	0.5782317409821224	0.06705693581873443
Facebook	Louvain	Random	0.9045249195225749	0.8786853395430391	0.8052424532165902	0.5724611973059331	0.37236137646294065	0.011963818958358984
Facebook	Louvain	Edge Jaccard JC	0.48403402858300393	0.3651460379898041	0.2647231473070433	0.14542491648597	0.06615101539927035	0.01507808997958444
Facebook	Louvain	Edge Local Search L-Spar	0.42567513206677404	0.34779321931478896	0.2686818747076462	0.06570211531004953	0.06602540918021237	0.06570211531004953
Facebook	Louvain	Clustering Coefficients	0.027057431657749993	0.021117406136713403	0.018534989230426 81	0.008101878837308995	0.006657045491188185	0.00265494282948249
Facebook	InfoMap	Edge Betweenness	0.538203170765926	0.7422483672954663	0.19994705082359895	0.5743790431794921	0.7391328930116096	0.15535304454043797
Facebook	InfoMap	Random	0.8967146312571734	0.7302115456709294	0.643330621403476	0.4328176192042061	0.7082439961145095	0.005612218183647099
Facebook	InfoMap	Edge Jaccard JC	0.548323353957553	0.27981834357185253	0.11991292244006696	0.057802622243655476	0.026132542692985233	0.0014710622821150685
Facebook	InfoMap	Edge Local Search L-Spar	0.6694704667080543	0.5540881917693584	0.1916402468031276	0.022934189714474758	0.023751837381374024	0.023751837381374024
Facebook	InfoMap	Clustering Coefficients	0.5459751951845913	0.21957283869644914	0.06199909279039507	0.021492670998236957	0.01525185671002377	0.0012668653977214819
Facebook	LPA	Edge Betweenness	0.6900099781229866	0.7144984460573799	0.49804844328892445	0.5506704136463123	0.6622062262256109	0.04330746649540818
Facebook	LPA	Random	0.4400600740113471	0.3885801701756929	0.2857144553475003	0.1771436534491125	0.15406138235517985	0.0018217215330985895
Facebook	LPA	Edge Jaccard JC	0.23029124027629602	0.16642677712082973	0.13302739639249878	0.07572002734858338	0.017041768725847273	0.00728105761059061
Facebook	LPA	Edge Local Search L-Spar	0.12807964148480655	0.0620380000327 6676	0.017336767624332634	0.009226810541251504	0.009226810541251504	0.009226810541251504
Facebook	LPA	Clustering Coefficients	0.21987858138258765	0.08556815907524754	0.04361709941114901	0.019593292083027434	0.01085110765584782	0.0016289884019114334

Table 3: Normalised Mutual Index (NMI) for Different Sparsifying Techniques

Network	Algorithm	Technique	Metric					
			Normalised Mutual Information (NMI)					
		Degree of Sparsification (%)	50	30	20	10	5	1
Amazon	Louvain	Edge Betweenness	0.9444939631784701	0.9110469374988062	0.8894966220176611	0.8555078874992165	0.8337576002259394	0.816204308134764
Amazon	Louvain	Random	0.9819774537439164	0.9354296178966223	0.8956775583036969	0.845328428114445	0.8269086826943726	0.8147457603975311
Amazon	Louvain	Edge Jaccard JC	0.9388172152292559	0.8956773433063665	0.86428651322174	0.8370587547606023	0.8235451274900741	0.8145747607702678
Amazon	Louvain	Edge Local Search L-Spar	0.9896436737756871	0.9518153163436417	0.9455747897555784	0.9455747897555784	0.9455747897555784	0.9455747897555784
Amazon	Louvain	Clustering Coefficients	0.6482471732509381	0.7162821410972223	0.7505089142146814	0.7853993164655863	0.7986467428735334	0.807899040301404
Amazon	InfoMap	Edge Betweenness	0.928252847056807	0.8869445500014788	0.8702740446262993	0.8407752356863685	0.8272026885368255	0.8151222791150288
Amazon	InfoMap	Random	0.9025758510287892	0.8901562635609885	0.8722550087132307	0.8386968746337249	0.8257332090648463	0.8143628741755055
Amazon	InfoMap	Edge Jaccard JC	0.8865229750612836	0.8717327348926273	0.8510688873476315	0.8323879809212384	0.822157113788331	0.8143175932284016
Amazon	InfoMap	Edge Local Search L-Spar	0.9037410686344515	0.8972224193422988	0.8982261211174132	0.8982261211174132	0.8982261211174132	0.8982261211174132
Amazon	InfoMap	Clustering Coefficients	0.8919493754559508	0.85970150949044	0.8447634746253097	0.8268754155732675	0.8195531437648517	0.8146359873705127
Amazon	LPA	Edge Betweenness	0.8705216855355602	0.8596290118341234	0.8508586811761775	0.8352754798497989	0.8292598659277444	0.8214352967849597
Amazon	LPA	Random	0.8919661571073709	0.8740602091479415	0.8568478270784856	0.8410378292339961	0.830816649427255	0.8219858936835909
Amazon	LPA	Edge Jaccard JC	0.8747236436414456	0.8600920780015948	0.8508571406190151	0.8386115207409642	0.8281034338973896	0.821666659081666
Amazon	LPA	Edge Local Search L-Spar	0.8922639960821123	0.8823339844201988	0.8816601564579328	0.8816601564579328	0.8816601564579328	0.8816601564579328
Amazon	LPA	Clustering Coefficients	0.8758655759159628	0.8581339734432715	0.8463658240819669	0.8313752152599416	0.826345660076481	0.8215617320605928
		Degree of Sparsification (%)	50	30	20	10	5	1
DBLP	Louvain	Edge Betweenness	0.9504288235694246	0.9187982728152281	0.8929163129167593	0.8459301918075732	0.8192220111724137	0.7996699322179449
DBLP	Louvain	Random	0.9667770137176271	0.9263544316646096	0.8880428068205198	0.8377495524844958	0.8149342142533165	0.8000034937974229
DBLP	Louvain	Edge Jaccard JC	0.8966346105429583	0.8746581017772277	0.8559777745970696	0.8310112308032889	0.8143939373489381	0.8004844486609816
DBLP	Louvain	Edge Local Search L-Spar	0.9527254417545277	0.9348128755023655	0.9316704008054792	0.9316704008054792	0.9316704008054792	0.9316704008054792
DBLP	Louvain	Clustering Coefficients	0.7044818490373401	0.7363531744789766	0.7566081682492101	0.77607149075975	0.78509310015602	0.7939965300146704
DBLP	Louvain	Effective Resistance Sampling	0.6384859380275986	0.6903769949233708	0.7265302942627797	0.7685261051346898	0.7850704777365723	0.796464096896343
DBLP	InfoMap	Edge Betweenness	0.9079689816739553	0.9042495173858551	0.8691552230350491	0.8351894274049193	0.8185651939267263	0.7948185398904334
DBLP	InfoMap	Random	0.896991824735255	0.8796160857031053	0.8599132792082862	0.8266515934107868	0.8072477603244632	0.7940420210904083
DBLP	InfoMap	Edge Jaccard JC	0.8600267598407042	0.8568012255005966	0.8397499072813902	0.8211369932792697	0.8070332930933083	0.7947339606513211
DBLP	InfoMap	Edge Local Search L-Spar	0.8939319618988014	0.8843328070628624	0.8912430269402053	0.8912430269402053	0.8912430269402053	0.8912430269402053
DBLP	InfoMap	Clustering Coefficients	0.8567035728962398	0.8389412543381499	0.8255135698553417	0.8096658769373918	0.8024974388689261	0.7942111014335907
DBLP	InfoMap	Effective Resistance Sampling	0.8904794350760923	0.8700593110126984	0.8497187883281545	0.8199227446743002	0.8024348225338547	0.7918544037156264
DBLP	LPA	Edge Betweenness	0.8722245187501378	0.8618554934416297	0.8536845422833345	0.8410473080273129	0.8354464967396369	0.8278203252501779
DBLP	LPA	Random	0.8828660057827167	0.8768497799539727	0.8623206436780695	0.8498778938458414	0.8377031408872793	0.8289952062300636
DBLP	LPA	Edge Jaccard JC	0.8559253344746682	0.855747748999679	0.8469701662860841	0.8397975493155198	0.837093043945366	0.8289359099215734
DBLP	LPA	Edge Local Search L-Spar	0.8880535952279391	0.8927129728986457	0.8901971509053086	0.8901971509053086	0.8901971509053086	0.8901971509053086
DBLP	LPA	Clustering Coefficients	0.8719066212365476	0.8571663671019599	0.8483793628776535	0.8384530130140408	0.8323667345550864	0.8286664105180844
DBLP	LPA	Effective Resistance Sampling	0.8915079963973033	0.8740581008905138	0.8626240584065212	0.8442142501617672	0.8337390912666988	0.8269146677960951
		Degree of Sparsification (%)	50	30	20	10	5	1
Email EU	Louvain	Edge Betweenness	0.42724835265874384	0.22494279528705952	0.20145876961492526	0.2492837673304421	0.3361175419629084	0.463480806200867
Email EU	Louvain	Random	0.7254223464563122	0.572365664045737	0.48870072371524004	0.40006133479972666	0.36917853476629675	0.4478691396368314
Email EU	Louvain	Edge Jaccard JC	0.5993971889644718	0.5938002119806703	0.568053955387778	0.5405938433067684	0.5215952947553679	0.4727727589464073
Email EU	Louvain	Edge Local Search L-Spar	0.7320490123517768	0.670411024616331	0.6264271794428706	0.5687067348736102	0.5687067348736103	0.5687067348736103
Email EU	Louvain	Clustering Coefficients	0.1567610702105313	0.19071979884095444	0.22905441885482578	0.3011678895809501	0.3664062019053673	0.44119018801640797
Email EU	InfoMap	Edge Betweenness	0.17492997869193772	0.7065765129653378	0.723733073449104	0.6726623274646625	0.6903381286870157	0.5200732783201802
Email EU	InfoMap	Random	0.824671604028161	0.7765780314164885	0.7052498398211393	0.6117384910558242	0.5750320376206784	0.5213074135392765
Email EU	InfoMap	Edge Jaccard JC	0.7577588158202578	0.6729739622919777	0.7129801900311056	0.693694719914861	0.5944638623784537	0.5176862408119394
Email EU	InfoMap	Edge Local Search L-Spar	0.3654112930370265	0.838766337826007	0.7635158756361584	0.6604673090627263	0.6604673090627263	0.6604673090627263
Email EU	InfoMap	Clustering Coefficients	0.7460196541977872	0.7119745449565501	0.7030160477452304	0.6719156567556035	0.604158755235936	0.5220889288665417
Email EU	LPA	Edge Betweenness	1.0	0.9223134002443141	0.6475510193055447	0.06102125392173808	0.05268186753694046	0.043916861781000441
Email EU	LPA	Random	0.33317508324043454	0.23523878456809555	0.15932393490545604	0.0979756728093113	0.047857051022328505	0.043719533113832619
Email EU	LPA	Edge Jaccard JC	0.08770212949707433	0.0751543947602262	0.06554417213739998	0.05535636490949204	0.047563126662099026	0.043575929219857312
Email EU	LPA	Edge Local Search L-Spar	0.0781694944709957	0.051099736375725985	0.04525984902326263	0.043182490692993465	0.043182490692993465	0.043182490692993465
Email EU	LPA	Clustering Coefficients	0.11634296467715954	0.07246223307534135	0.06281139722866888	0.05429948578063651	0.04853577916000249	0.043715956443200324
		Degree of Sparsification (%)	50	30	20	10	5	1
Facebook	Louvain	Edge Betweenness	0.9209237136383327	0.859920063352129	0.8237301770757645	0.7832538386833904	0.7389026436902781	0.47842351056061067
Facebook	Louvain	Random	0.916492177863189	0.8783506248400579	0.8171800253508567	0.6937827570042214	0.5989147673568086	0.47572090358354546
Facebook	Louvain	Edge Jaccard JC	0.7524092281944504	0.6821693422752613	0.6164870667938884	0.5413229972029453	0.4976312929180678	0.4708205993745063
Facebook	Louvain	Edge Local Search L-Spar	0.7476833299568821	0.7134087864179379	0.6598978421012989	0.576302562292105	0.5763504675580158	0.576302562292105
Facebook	Louvain	Clustering Coefficients	0.24547162864502464	0.3370860668225476	0.4251322200841834	0.44553678964572063	0.4493089145835036	0.4612664638620029
Facebook	InfoMap	Edge Betweenness	0.6539170351654019	0.7639801711608049	0.3781432376502583	0.7321624346495422	0.7188380369339245	0.3758563074409872
Facebook	InfoMap	Random	0.8418386203276057	0.7114971066185993	0.6393207996227825	0.5607193784148239	0.5251285422125613	0.346639063813332
Facebook	InfoMap	Edge Jaccard JC	0.6265424101800806	0.553857973231861	0.4626085173819689	0.3967755314082553	0.3620915717612197	0.3357884788040875
Facebook	InfoMap	Edge Local Search L-Spar	0.694689526768193	0.6143634192441337	0.5159697680389482	0.42365161799707135	0.4241921961916266	0.4241921961916266
Facebook	InfoMap	Clustering Coefficients	0.5778970523469185	0.4499753351370318	0.37340453161816206	0.35097499591464104	0.346832678887947	0.3349388622553799
Facebook	LPA	Edge Betweenness	0.7453645959533861	0.7655843990547844	0.7158282508349414	0.7114904415447949	0.7177967322444637	0.4683463484401772
Facebook	LPA	Random	0.721040747294837	0.6885286894334326	0.6348885544318585	0.5445868513858307	0.5034047365190857	0.4535182807326248
Facebook	LPA	Edge Jaccard JC	0.6369172011881598	0.5828454555117867	0.5558659490194565	0.5011324836605571	0.4630812443934718	0.44942681968627046
Facebook	LPA	Edge Local Search L-Spar	0.6079189350404239	0.5476116427902883	0.511313083186543	0.4966426596395599	0.4966426596395599	0.4966426596395599
Facebook	LPA	Clustering Coefficients	0.5977581105651232	0.5280783831439093	0.4791418541970022	0.45718877991434126	0.4513358563811198	0.4464780287755262