Centrality Clustering-Based Sampling for Big Data Visualization

Tam Thanh Nguyen
School of Business/IT
James Cook University Australia
Singapore Campus, Singapore
thanhtam.nguyen@my.jcu.edu.au

Insu Song
School of Business/IT
James Cook University Australia
Singapore Campus, Singapore
insu.song@jcu.edu.au

Abstract—**Information visualization is essential for improving effectiveness and efficiency of data exploration and knowledge discovery. Therefore, visualization has been used in a wide range of fields from biology, medicine, criminal activity analysis to business and education. Information visualization has become more important than ever as the amount of data being generated has increased dramatically in recent years. One of the major difficulties of information visualization is performance, and this is even more critical when visualizing big data. One potential solution to this challenge is data sampling while maintaining fidelity of visual representation. In this paper, we propose two new centrality clustering-based sampling approaches that apply centrality measures on clusters of data points in order to make more informed sampling than random sampling approaches. We evaluate the new methods on graph data sets. The results show that the new methods significantly outperform existing data sampling methods in term of perceived differences and their ability to preserve essential visual information. Moreover computational complexity is comparable or even better than simple random sampling methods.**

Keywords— big data, big data visualization, data sampling, clustering algorithm, cluster sampling, large data sets

I. Introduction

Visualization has been studied in various researches and been considered as a useful tool for effective and efficient improvement in the process of data exploration and discovery [1, 2]. Keller and Tergan [3] stated that humans could think and understand best with the help of cognitive tools since the speed of pattern recognition from visual display was faster and more effective than that of accessing and processing data in the brain. Sharing the same opinion, [1] contended that most cognitive processes were done as some certain type of interaction with cognitive tools, such as papers and pencils to information visualization. He claimed that visualization was playing an increasingly critical role in boosting people's cognition.

One of the key advantages of data visualization is that it could interpret a large amount of data and present it in a way that users can quickly perceive the insights into that data [1]. Several benefits that data visualization can offer were discussed, including (1) enabling comprehension of large data sets, (2) allowing unexpected patterns and properties to be easily recognized, (3) highlighting any issues existing in the data, and (4) assisting the formation of hypothesis from the visual display.

Given these remarkable advantages, visualization has been employed in many fields, including biology, social network analysis, education and business. In the age where data is being created, the value of visualization has become more crucial than ever in order to obtain thorough insights into the data. Turner, et al. [4] stated that the digital world by 2020 will comprise of the same amount of digital bits as number of stars in the universe. The world's digital data is experiencing a twofold growth every two years and is projected to increase from 4.4 zettabytes in 2013 to about 44 zettabytes by 2020. In addition, people can know very little about data if it is not characterized. Yet, in 2013, more than 20% of the entire data of the digital world could be useful if it was tagged; while less than 5% of that data was analyzed [4]. Accordingly, big data has required approaches to presenting the massive digital data so that it is easier for users to understand and use. That was when visualization came into play and created significant impacts.

However, opportunities come fraught with challenges. The big data visualization process takes too long to complete because of the huge amount of data to be visualized. As mentioned in [5], with the amount of data being collected every day, due to its enormous size, the speed of data visualization gets affected. One potential solution to this challenge is sampling while maintaining fidelity of the visual representation. In this paper, we propose two novel intelligent sampling approaches which utilize centrality measures and clustering methods in order to make informed sampling. Within the scope of this research, the new approaches are evaluated under a thorough process of visualization. The test data set is put through several phases: acquire, parse, filter, mine, and represent.

In particular, we describe two novel centrality clustering-based sampling methods: Centrality Ranked Sampling (CRS) and Cluster Walk Sampling (CWS). These methods perform clustering on the data, and then perform sampling on the clusters using eigenvector centrality measures of the data points. We show that the new sampling methods significantly outperform existing sampling methods, such as random sampling and random walk sampling, in terms of scalability and ability to preserve essential information for data visualization.

978-1-5090-0620-5/16/$31.00 ©2016 IEEE

The paper is organized as follows: In Section II, we describe related works on data visualization, challenges in big data visualization, data sampling approaches, graph clustering and graph sampling methods. In Section III, we describe our new sampling methods and data sets that are used for evaluation. In Section IV, we compare the new sampling methods with existing sampling methods. In Section V, we conclude the paper with remarks.

II. Background

A. Data Visualization

Data visualization has been used from long time ago [6]. During medieval times, maps were used to facilitate navigation and exploration. Its evolution then, has gone through several stages such as measurement and theory, new graphic forms like abstract graphs, modern graphics like time series graphs, statistical graphics; and now high dimension interactive visualization. A great variety of methods and techniques have been generated and developed to contribute to the development of data visualization. The key cause for the birth and improvement of this area was the desire to perceive data and relationships from different and novel viewpoints.

To define what data visualization is, many researchers have proposed various ideas and methods based on theoretical, technological and empirical studies. Due to many available forms of visualization and technologies, and analysis methods, many definitions have been proposed. Therefore, no uniformly acceptable definition exists [7].

Card, et al. [2] stated that Infovis (Information visualization) is about making use of the dynamic, interactive, inexpensive medium of graphical computers to offer new external supports that enhance human cognitive abilities. In another article, Wills [8] demonstrated that people keen on data would be willing to explore data in its visualized form. He defined information visualization as a process that accepted data as the input and generated the visual illustration of those data as the output. Representation had to be readily understood by the users.

Moreover, as mentioned in the study titled as "A taxonomy of clutter reduction for information visualization", information visualization can be considered a tool for gaining insight into data through a visual representation [9]. Fayyad, et al. [10] stated that there has always been a desire for understanding structures, patterns and relationships in data, and visualization supports this by providing data in a variety of forms with different types of interaction.

Ware [1] summarized the advantages of visualization as follows: (1) enabling the comprehension of large data sets, (2) allowing unexpected ways of perceiving information, (3) facilitating identification of errors in data, (4) allowing cognizance of linking patterns, (5) facilitating the formation of hypotheses.

B. Performance Issue in Visualization for Big Data

For big data applications, it is particularly difficult to conduct data visualization because of the large size and high dimension of big data. Regardless of the many benefits provided by big data visualization, there is a major challenge: performance. The data visualization process consists of transformation and representation processes. The former deals with the conversion of data into the graphical primitives, whereas the latter handles and stores all the outputs of these processes [11]. In the meantime, the first key feature of big data is volume, and it is huge. That results in a decreased performance rate in its rendering due to the size and scale of the data [5]. Typically, with the amount of data being collected day by day, the speed of the visual rendering of the data slows down due to its sheer size. In most cases the average data size is in gigabytes or terabytes. Simply loading all of this data into memory can tax it greatly.

The speedy expansion of social networks has given researchers opportunities to study social processes, interactions, and relationships [12-16]. Furthermore mobile phone peneration rate in the world has given opporunities to provide cost effective finance, education, and health services [17-21]. However, opportunities came with challenges due to the large amount of data, which included "data acquisition bottleneck" and "information analysis complexity" [7, 16, 22]. Likewise, Leskovec and Faloutsos [23] pointed out that an Online Social Network (OSN) graph could contain millions of nodes and edges, which made it unfeasible to store the whole graph in the computer's memory. They contended that even when it was possible, the process would be extremely time consuming.

This limitation in big data visualization necessitates data reduction since it is not always possible to work with data in its full form. Even when it is possible, the process is inconvenient. In this paper, we propose a strategy to reduce the amount of data to visualize based on a sampling approach.

C. Sampling Approaches

Many sampling approaches have been developed to reduce the amount of data to be analyzed and visualized. For very large data sets, probability sampling methods have been utilized, including Simple Random Sampling, Systematic Sampling, Sequential Random Sampling, Stratified Sampling and Cluster Sampling.

Simple Random Sampling (SRS) is a sampling method where all observations in the population have the same probability to be chosen as the sample. However, this method may not be representative and may include outliers and unimportant structures in the sample [24].

Systematic Sampling chooses observations for the sample based on a constant interval [25]. This approach, when planned carefully, can provide more information per unit cost than SRS and guarantee that the sample will be uniformly spread over the population. Yet, the sample generated with this approach may not be representative if the population is in a periodic order [25].

Stratified Sampling [25] divides the population into strata, and observations are taken proportionally from the strata to construct the final sample. There is one drawback of this method – prior knowledge and pre-processing are required [24].

Cluster Sampling is stated to be simple, inexpensive and relatively effective for implementation and evaluation; however, it possesses one key drawback. The objects within the same cluster have more similar characteristics compared to those in different clusters. Consequently, the observations in a selected cluster provide less information than observations in not selected ones.

D. *Graph Sampling*

Networks and social graphs can also be counted as a type of very large data sets. Sampling approaches for this type of data usually include Random Walk sampling [26], Snowball sampling [23], Forest Fire sampling [12, 23] and Metropolis-Hastings Random Walk [27].

In Random Walk sampling, an initial node is chosen at random and then one of its neighboring nodes is selected randomly or according to some weight. The process continues till the number of nodes selected reaches a specific sample size [26]. Snowball sampling is similar to Random Walk at the first step, which is randomly picking an initial node. However, a fixed number of nodes are added into the sample. These nodes directly connect with the initial node via its outgoing edges. Random Walk and Snowball sampling are generally inexpensive, simple and useful for exploring the graphs' topology; but they are highly biased to high degree areas [23].

Forest Fire sampling is a probabilistic version of Snowball Sampling [26]. One key feature of this approach is its ability to avoid re-visiting the visited nodes [12]. This sampling method is useful to capture graph topology but not for information content and social context [12].

Another sampling technique for large graphs is Metropolis-Hastings Random Walk. In this approach, the transition probabilities are modified in order that the samples generated conform to the desired uniform distribution [28]. Metropolis-Hastings Random Walk can help to correct the bias to a high degree [28], but its application is more limited than that of Random Walk, since the degree of neighboring nodes must be known beforehand [26].

E. *Graph Clustering*

As mentioned in [29], a good clustering method should generate (1) clusters with intra-cluster density greater than that of the whole graph and (2) inter-cluster density is smaller than that of the graph. Dealing with the data set used, the clustering process can be understood as community detection. There are several algorithms to identify communities within a network graph: hierarchical clustering, Girvan-Newman [30] which identifies edges that connect two different communities and remove them to obtain only communities, and modularity optimization [31] which perform iterative search over divisions within the network for ones with high modularity.

All in all, previous approaches for large dataset sampling are either simple and inexpensive, or complicated. With simple methods, the sampling process might be fast and easy to implement; however, there is no guarantee that the sample is representative of the original data. In contrast, with more complicated approaches, such as Metropolis-Hasting Random sampling, implementation is longer and more complex, especially for big data with huge amounts of observations to process.

III. Methodology

A. *Dataset*

We illustrate our centrality clustering-based sampling methods using a publicly available graph dataset. The data set is also used to evaluate the performance of the methods in the next section. The data set is retrieved from snap.stanford.edu (Stanford Network Analysis Platform). It was collected by conducting a survey on participants who used Facebook. The data contains link information between Facebook users. User IDs were anonymized by using natural numbers (from 0) to replace the real Facebook internal IDs. Users are represented as nodes in the graph. There exists a link between two nodes if the two users are friends on Facebook. This data therefore describes an undirected network graph as shown in Fig. 1. Table 1 shows the statistics of the undirected graph.

Fig. 1. Facebook friend graph. Colors are used to show clusters in the graph.

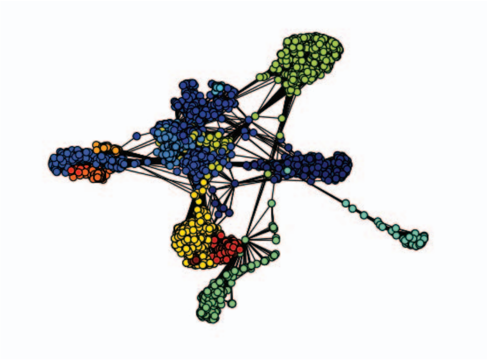


TABLE I. Description of Data Set Used

Graph Features	Size
Number of nodes	4,039
Number of edges	88,234
Average degree	43.6910
Diameter (longest shortest path)	8
Average clustering coefficient	0.6055
Number of nodes	4,039

[a] Source: https://snap.stanford.edu/data/egonets-Facebook.html

B. Clustering-based Sampling

Fig. 2. Overview of centrality clustering-based sampling.

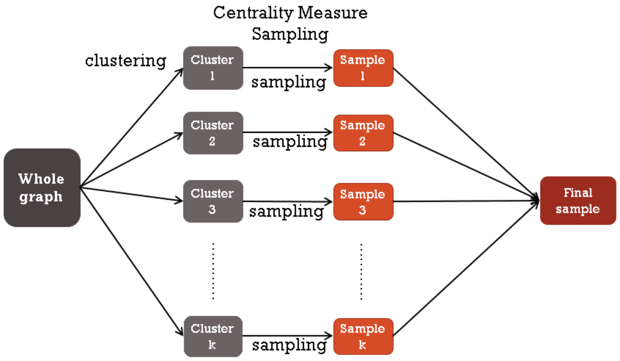


Fig. 3. Illustration of five visualization approaches.

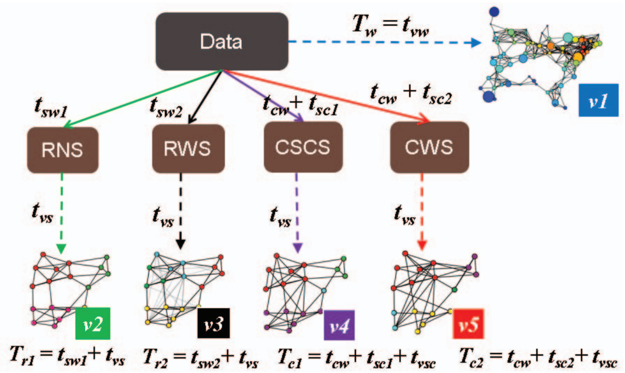


Fig. 2 illustrates the overall approach of the centrality clustering-based sampling methods. Given the original dataset, we first perform clustering on the dataset to generate clusters, where each cluster consists of data points that are similar. The clustering step is to form distinct clusters from the original graph: global clustering. In our experiment, we use Louvain method [31] for clustering, which clusters a graph by optimizing local communities until global modularity cannot be changed to a better state.

Next, from the generated clusters, we take samples from each of the clusters. We develop two new sampling methods for graph data: Centrality Ranked Sampling (CRS) and Cluster Walk Sampling (CWS). These sampling methods are described in the next sub-section in more details. We then combine the samples from the clusters to construct the final sample of the dataset.

Fig. 3 illustrates 5 methods of data visualization: v1: directly visualizing without sampling, v2: using random sampling, v3: using random walk sampling, v4: CRS sampling, and v5: using CWS sampling.

C. Cluster Sampling Methods

Centrality Ranked Sampling (CRS) samples clusters based on eigenvector centrality which reflects more realistic centrality of vertices than simple degree of centrality [32]. A degree of centrality of a vertex is the number of edges (connections) to other vertices, whereas eigenvector centrality incorporates the degree of importance of neighboring vertices: eigenvector centrality computes the approximate importance of each node in the graph. The eigenvector centrality v_i for vertex i can be obtained as follows:

$$v_i = \frac{1}{\lambda} \sum_i A_{ij} v_j$$

where A is the adjacency matrix of the graph. The eigenvector centrality v_i satisfies the following condition:

$$Av = \lambda v$$

where λ is the largest eigenvalue and v is an eigenvector. Due to Perron–Frobenius theorem, λ is the largest eigenvalue of the adjacency matrix with the corresponding eigenvector [32].

After obtaining the eigenvector centrality of the graph, we sort the vertices in each cluster in descending order of eigenvector centrality of the vertices. Some top portions of the vertices in the clusters are then sampled to form the final sample.

Cluster Walk Sampling (CWS) samples clusters by performing random walk [26] on the clusters. The vertex with the highest eigenvector centrality value is selected as the starting node. CWS then randomly visits a neighboring vertex and so on until it reaches the set number of vertices to be sampled. The random walk algorithm is implemented to avoid re-visiting any visited nodes.

IV. Experiments

The proposed clustering-based sampling methods are evaluated using Facebook dataset. We compare our sampling algorithms with conventional random sampling approaches for their performance in terms of efficiency and their ability to keep essential information for visualization. We measure both visually perceived differences and information fidelity metrics. For visually perceived differences, we inspect whether the clusters that can be formed from the original graph can still be identifiable after sampling. To do this, we cluster the resulting data sets and use color coding to inspect the results visually.

For information fidelity metrics, we adopt the set of quantitative performance measures (characteristics of sample graphs) defined by Albert [33]. The metrics comprise of diameter, average path length, clustering coefficient, and degree distributions. Diameter of a graph is defined as the maximal distance between two nodes in the graph. Average path length is the average number of steps that need to be taken for all possible pairs in the graph to reach each other. Clustering coefficient is a concept used to measure the degree to which nodes in the graph have a tendency to cluster together. Clustering coefficient of the whole graph is the average of all single nodes' clustering coefficient. Degree distribution is the spread in the node degrees (number of edges going through the node) across the network graph.

Finally, we measure efficiency of the approaches: running times. The time needed for sampling and visualization is recorded to compare computational complexity of the approaches. Without sampling, the time Tw that it takes to visualize the whole dataset is simply the time t_{vw} for visualizing the whole dataset:

$$Tw = t_{vw}$$

For random sampling approaches, the time Tr includes the time t_{sw} for sampling from the whole dataset and the time t_{vs} for visualizing the sampled dataset:

$$Tr = t_{sw} + t_{vs}$$

For random sampling, we use random node sampling and random walk sampling described in [34]. In random node sampling (RNS), nodes (vertices) are randomly chosen from the whole graph, and the sample is induced from this set of nodes. In Random Walk sampling (RWS), one node is chosen randomly as a starting point, then random walk is simulated on the graph. In our experiment, Random Walk sampling does not take visited nodes into the sample since this network graph is undirected and once a node is put into the sample, it stays in the sample.

For cluster-based sampling approaches, the time Tc includes the time t_{cw} for clustering the whole dataset, the time t_{sc} for sampling from the clusters, and the time t_{vsc} for visualizing the sampled clusters:

$$Tc = t_{cw} + t_{sc} + t_{vsc}$$

For centrality clustering-based sampling, we use CRS and CWS for sampling. Fig 3 illustrates the above 5 visualization processes.

A. Experimental setup

In order to determine the performances of the sample generation processes and the qualities of samples generated, we perform comparison between the visualization of the entire data set, RNS, RWS, CRS, and CRW samples.

For the whole graph, we perform clustering to identify distinct groups and visualize those groups with different colors for visual comparison.

As for random sampling, we experiment Random Node Sampling (RNS) and Random Walk Sampling (RWS). Stumpf, et al. [34] showed that random node sampling does not retain power-law degree distribution. However, in the experiment, we simply want to find a good sampling approach for the given graph which generates good results that meet our evaluation criteria. Leskovec and Faloutsos [23] used Random Node, as well as Random Walk sampling in their research. With Random Node sampling, a set of nodes are randomly chosen from the whole graph, and the sample is induced from this set of nodes. With regard to Random Walk sampling, one node is chosen randomly as a starting point, then random walk is simulated on the graph. In our experiment, Random Walk sampling does not take visited nodes into the sample since this network graph is undirected and once a node is put into the sample, it stays in the sample.

B. Experimental Results

1) Visually Perceived Differences

Fig. 4. Visualization results of the samples generated using Random Node Sampling (RNS), Random Walk Sampling (RWS), Centrality Ranked Sampling (CRS), and Cluster Walk Sampling (CWS)

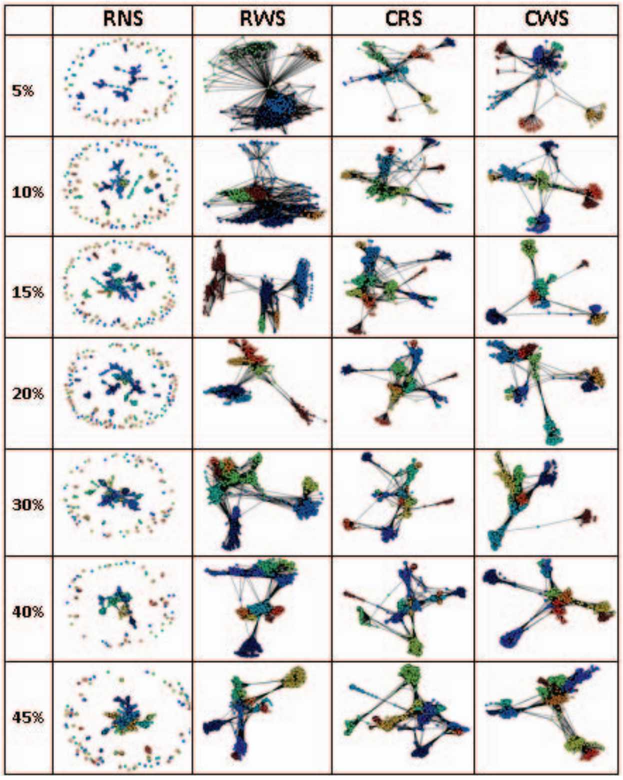


Fig. 4 shows the visualization results of the samples generated using RNS, RWS, CRS, and CWS. We generate samples from the population graph using these sampling approaches at different sample sizes (5%, 10%, 15%, 20%, 30%, 40%, and 45%).

As can be seen, the samples generated using RNS are sparse and not able to preserve any information even at 45%. The samples produced using RWS, CRS and CWS show tight connections among nodes and clusters. Visualization of samples created using RWS have fewer distintive clusters than CRS and CWS. CRS generates samples with more distingtive clusters than any other sampling approaches.

The original entire graph has 16 distinctive clusters. This can be used to quantitively measure how much information is preserved after sampling. Table II shows the average number of clusters of generated samples at different sampling sizes. It shows than CRW and CWS are far superior than all other sampling approaches in preserving information for data visualization.

TABLE II. AVERAGE NUMBER OF CLUSTERS OF SAMPLED DATA SETS

	5%	10%	15%	20%	30%	40%
RNS	83	112	112	108	84	74
RWS	5	6	7	7	9	9
CRS	7	11	12	13	13	13
CWS	8	9	9	10	9	12

Fig. 5. Average diameters of clusters of data sets. CRW and CWS outperform RNS and RWS.

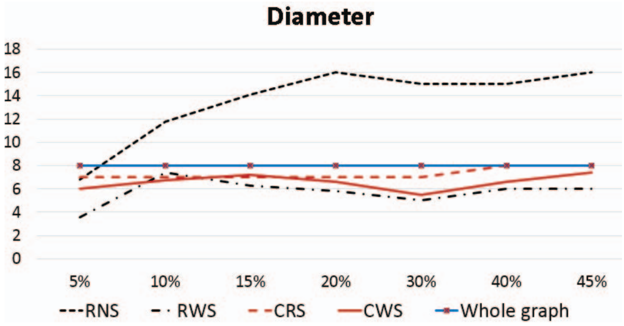
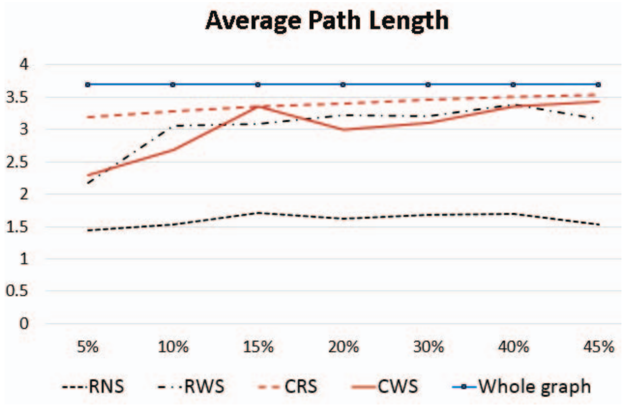


Fig. 6. Average path lengths of data sets. CRS outperforms all other methods.



C. Information Fidelity Metrics

As shown in Fig. 5, the average diameter of samples generated by CRS is most similar to the original graph (the whole dataset), followed by CWS and RWS. RNS is least similar to the original graph in terms of diameter indicator.

Fig. 6 shows average path lengths over different sampling sizes. CRS again generates samples with the most similar average path length to that of the original graph. RNS creates samples with the lowest average path length among all approaches.

Fig. 7 shows the average clustering coefficients of the samples. Samples generated by CWS are very similar to the original one. CWS and CRS also have good results for this indicator.

Fig. 7. Average clustering coefficient of the data sets.

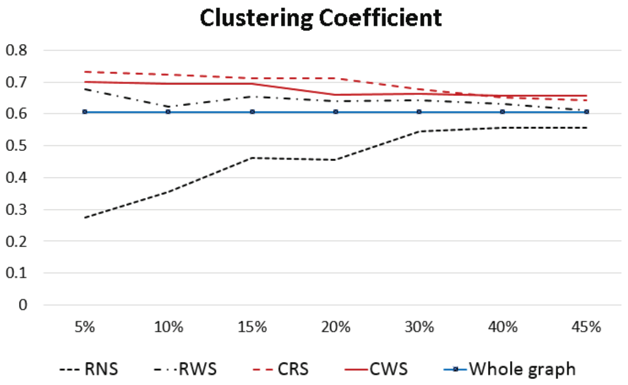
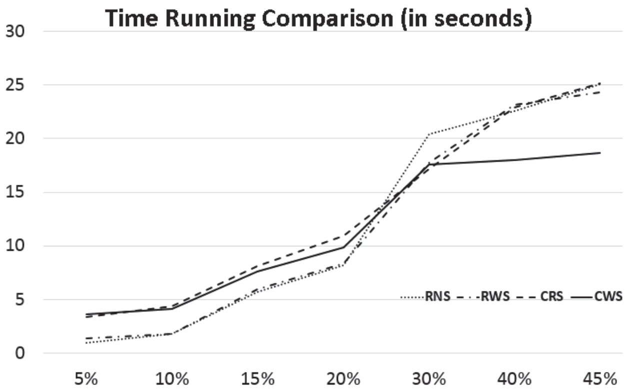


Fig. 8. Run-time of 4 sampling approaches.



The whole graph has a long tail indicating, that the original network has few nodes with a much higher degree of distribution than most other nodes. The majority of nodes has degrees ranging from low to medium. Samples generated using CWS and RWS have similar patterns to the original graph in terms of the degree of distribution. The CWS, CRS and RNS gets better when the sample size increases to more than 35%.

Fig. 8 shows the run-times of the 4 sampling approaches. At sample size less than 30%, RNS and RWS performs only slightly better than CRS and CWS. With sample size greater than 30%, CWS is faster than other approaches. This indicates that CRS and CWS are scalable for large data sets.

D. Discussion

RNS generates samples with scattered nodes and has the worst measures for information fidelity metrics compared to other approaches. CWS, RWS and CRS generate better quality samples for the original network graph. Average path length and diameter of samples using CRS and CWS are more similar than RWS; while RWS performs better with respect to Clustering Coefficient indicator. However, the degree distribution of samples generated by CWS are slightly better than that of samples created using RWS and CRS, which indicates that the structure of samples generated by CWS are more similar to the original network than those generated by RWS and CRS.

V. Conclusions

The proposed method for sampling from big data sets using clustering helps retain the characteristics of the original dataset. This paper introduces two new improved methods of sampling using centrality measures of clusters. With CRS, nodes with high eigenvector centrality values are chosen to put into the sample in order to maintain the structure of the original graph. CWS performs random walk on clusters with the starting nodes being those vertices with the highest eigenvector centrality values in the clusters, generating samples with the most similar degree distribution to that of the original graph. Moreover, at sample size of 30% and greater, CWS even outperforms all other sampling approaches in speed.

Results of this paper show that CWS can be used to sample from undirected unweighted networks, for instance friendship networks, reducing the data to be visualized by 70% while maintaining the structure of the original data, with lower running time than Random Sampling approaches.

References

[1] C. Ware, *Information visualization: perception for design*: Elsevier, 2012.

[2] S. K. Card, J. D. Mackinlay, and B. Shneiderman, *Readings in information visualization: using vision to think*: Morgan Kaufmann, 1999.

[3] T. Keller and S.-O. Tergan, *Knowledge and information visualization*: Springer, 2005.

[4] V. Turner, J. F. Gantz, D. Reinsel, and S. Minton, "The digital universe of opportunities: Rich data and the increasing value of the internet of things," *International Data Corporation, White Paper, IDC_1672,* 2014.

[5] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov, "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration," *Briefings in bioinformatics,* p. bbs017, 2012.

[6] M. Friendly, "A brief history of data visualization," in *Handbook of data visualization*, ed: Springer, 2008, pp. 15-56.

[7] N. T. Tam and I. Song, "Big Data Visualization," in *Information Science and Applications (ICISA) 2016*, ed: Springer, 2016, pp. 399-408.

[8] G. Wills, "Visualization toolkit software," *Wiley Interdisciplinary Reviews: Computational Statistics,* vol. 4, pp. 474-481, 2012.

[9] G. Ellis and A. Dix, "A taxonomy of clutter reduction for information visualisation," *Visualization and Computer Graphics, IEEE Transactions on,* vol. 13, pp. 1216-1223, 2007.

[10] U. M. Fayyad, A. Wierse, and G. G. Grinstein, *Information visualization in data mining and knowledge discovery*: Morgan Kaufmann, 2002.

[11] E. H.-h. Chi, *A framework for visualizing information* vol. 1: Springer Science & Business Media, 2002.

[12] M. De Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher, "How does the data sampling strategy impact the discovery of information diffusion in social media?," *ICWSM,* vol. 10, pp. 34-41, 2010.

[13] I. Song, D. Dillon, T. J. Goh, and M. Sung, "A health social network recommender system," in *Agents in Principle, Agents in Practice*, ed: Springer, 2011, pp. 361-372.

[14] I. Song and N. V. Marsh, "Anonymous indexing of health conditions for a similarity measure," *Information Technology in Biomedicine, IEEE Transactions on,* vol. 16, pp. 737-744, 2012.

[15] J. Vong and I. Song, *Emerging Technologies for Emerging Markets* vol. 11: Springer, 2015.

[16] M. Lech, I. Song, P. Yellowlees, and J. Diederich, *Mental Health Informatics* vol. 491: Springer, 2014.

[17] I. Song, "Diagnosis of pneumonia from sounds collected using low cost cell phones," in *Neural Networks (IJCNN), 2015 International Joint Conference on*, 2015, pp. 1-8.

[18] I. Song, "Gaussian Hamming Distance," in *Neural Information Processing*, 2015, pp. 233-240.

[19] I. Song and J. Vong, "Mobile Collaborative Experiential Learning (MCEL): Personalized Formative Assessment," in *IT Convergence and Security (ICITCS), 2013 International Conference on*, 2013, pp. 1-4.

[20] I. Song and J. Vong, "Affective core-banking services for microfinance," in *Computer and Information Science*, ed: Springer, 2013, pp. 91-102.

[21] I. Song and J. Vong, "Mobile Core-Banking Server: Cashless, Branchless and Wireless Retail Banking for the Mass Market," in *IT Convergence and Security (ICITCS), 2013 International Conference on*, 2013, pp. 1-4.

[22] S. Chandrasekaran and I. Song, "Sustainability of Big Data Servers Under Rapid Changes of Technology," in *Information Science and Applications (ICISA) 2016*, ed: Springer, 2016, pp. 149-159.

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

[24] Z. Liu, B. Jiang, and J. Heer, "imMens: Real - time Visual Querying of Big Data," in *Computer Graphics Forum*, 2013, pp. 421-430.

[25] S. Lohr, *Sampling: design and analysis*: Cengage Learning, 2009.

[26] P. Hu and W. C. Lau, "A survey and taxonomy of graph sampling," *arXiv preprint arXiv:1308.5865,* 2013.

[27] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The journal of chemical physics,* vol. 21, pp. 1087-1092, 1953.

[28] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in Facebook: A case study of unbiased sampling of OSNs," in *INFOCOM, 2010 Proceedings IEEE*, 2010, pp. 1-9.

[29] M. E. Newman, "Detecting community structure in networks," *The European Physical Journal B-Condensed Matter and Complex Systems,* vol. 38, pp. 321-330, 2004.

[30] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the national academy of sciences,* vol. 99, pp. 7821-7826, 2002.

[31] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment,* vol. 2008, p. P10008, 2008.

[32] M. E. Newman, "The mathematics of networks," *The new palgrave encyclopedia of economics,* vol. 2, pp. 1-12, 2008.

[33] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics,* vol. 74, p. 47, 2002.

[34] M. P. Stumpf, C. Wiuf, and R. M. May, "Subnets of scale-free networks are not scale-free: sampling properties of networks," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 102, pp. 4221-4224, 2005.