IJWR

A Hybrid Seed Node Selection and No-Retracing Random Walk in Page Rank Algorithm

Azam Bastanfard*, Ali Kheradbeygi Moghadam, Ali Fallahi RahmatAbadi
Department of Computer Engineering, Karaj Branch, Islamic Azad University, Karaj, Iran;
bastanfard@kiau.ac.ir*, info@kheradbeygi.ir, ali.fallahi@kiau.ac.ir

ABSTRACT

The random walk technique, which has a reputation for excellent performance, is one method for complex networks sampling. However, reducing the input data size is still a considerable topic to increase the efficiency and speed of this algorithm. The two approaches discussed in this paper, the no-retracing and the seed node selection algorithms, inspired the development of random walk technique. The Google PageRank method is integrated with these different approaches. Input data size is decreased while critical nodes are preserved. A real database was used for this sampling. Significant sample characteristics were also covered, including average clustering coefficient, sampling effectiveness, degree distribution, and average degree. The no-retracing method, for example, performs better. The efficiency increases even further when the no-retracing technique is combined with the Google PageRank. When choosing between public transportation and aircraft, for example, these algorithms might be used since time is crucial. Additionally, these algorithms are more energy-efficient methods that were looked at.

Keywords: ***Random Walking, Network Sampling, Page-Rank Algorithm, Clustering, Markov Chain.***

1. Introduction

Political polling has influenced how sampling has changed in the social sciences. The American publication Literary Digest sent out postcards to readers whose names were pulled from telephone directories and car owner records in 1920. The magazine editors questioned which one of the candidates for the presidential election the audience would choose to support, in the postcards. Literary Digest gathered all the information and accurately projected who would prevail in the presidential contest. During the next presidential election period, the journal expanded its polling, and its predictions for the presidential election in 1924, 1928, and 1932 became trustworthy. In 1936, the magazine surveyed ten million individuals, receiving two million replies [1]. Alf London received 57 percent of the vote, while Roosevelt received 43 percent. However, the outcomes of the presidential election were substantially different, with Roosevelt winning 61 percent of the vote [2]. The sampling approach was where the Digest Literary editors went wrong. They chose a sample of wealthy people consisting of phone and private car owners [3].

Sampling approaches are generally classified as Probability or Non-Probability with Biased or Unbiased sampling methods depending on the sampling technique [4]. Each strategy outperforms the others in a number of features, but there is no method excellent from the others in every way. Finding a sample that is the most similar to the leading network is the most critical issue. Among these methods, navigation and search algorithms, as well as the evaluation of nodes in the network, have been used, and the combination of these methods has occasionally yielded positive results [5].

In research studies looking at the growth of social networks and cyberspace, such networks have been examined, and their prevalence among various demographics has received a lot of attention [6, 7]. Because of the expanding tendency to evaluate networks as large dynamic complex social graphs., much research has gone into identifying social networks [8]. In other words, there is a network everywhere information is transferred. Researchers have focused on social networks to accomplish outcomes and identify specific cases. To conduct their studies, each of these researchers requires a sample of these networks. The network models appropriate for the actual network are regarded as sampling due to their extensive scope and restricted admission to genuine networks [9]. The absence of the entire network significantly impacts social network sampling approaches. Several approaches for sampling from social networks have been presented based on these methodologies. According to the above, the statistical method must be used to sample from networks, which must be examined first. On the other side, the massive number of nodes and their connections negatively affect the network and network analysis. As a result, developing a practical approach for sampling social networks is critical [10, 11].

Computer networks and multidisciplinary sciences have gotten a lot of attention in recent years. Examples include but are not limited to mathematics, statistics, physics, computer science, etc [12, 13]. As participatory software, social networks will rely on their members. The concept of a social network is meaningless without users [14]. Additionally, some companies utilize social media for client engagement and marketing initiatives [15, 16].

Watts and Strogatz [17], Barabasi and Albert [18], and Erdos Renyi's [19] models in order to conceptualize and

doi ijwr.2022.344504.1115. 20.1001.1.26454335.2022.5.1.6.9
Citation. A. Bastanfard, A. Kheradbeygi Moghadam, and A. Fallahi RahmatAbadi, " A Hybrid Seed Node Selection and No-Retracing Random Walk in Page Rank Algorithm," *International Journal of Web Research*, vol.5, no.1, pp. 42-49, 2022.

*Coressponding Author
Article History: Received:27 May 2022; Revised: 22 July 2022; Accepted: 13 August 2022

investigate the real structure [20]. These artificial networks may be used to study complex networks' internal structure and characteristics. However, a portion of the network can be evaluated more closely to examine the whole network. Finding the best method to sample the subnetworks to reflect the entire network and how to extract a small but helpful sample, especially when the graph is vast and has many nodes and links, is essential in this situation.

There have been various types of research on sampling, including ones looking at how web spiders sample the Internet (www) [21]. The method of searching for online examples was also investigated. Crawlers take a portion of the data from publicly accessible sources, such as open graphs [22]. In some circumstances, the topological properties of the networks that were sampled may be quite different from those of the leading network. In actuality, our world is a networked one. Other network types, such as standard and randomized, do not possess the qualities these networks have. The term "Complex network" refers to the fact that most of these networks have a complicated structure.

Based on what has been mentioned, sampling in complex networks needs more hypothetical and experimental research and investigations. include such methods as random sampling and sampling that is reliant on connection [23]. In this context, it is essential to highlight the following factors:

1. The tiny scope and straightforward design of the structure enable more satisfactory sampling.

2. The main network values are derived by performing controlled steps on the average clustering coefficient as well as the mean degree of the sampled subset.

3. Both rankings and subcategories have a more comprehensive degree of separation.

Both theoretically and practically, there have been a lot of studies done on the random walk, and many scoring detection approaches were developed based on the classic random walk method [24, 25].

1.1 Markov Chain

The Markov chain was named after the Russian scientist. Markov is a mathematical system that enables transitions between states regardless of the beginning state and despite the existence of several states. Examples include but are not limited to math, data science, natural sciences such as physics, software engineering, etc [26]. The Markov chain is a memoryless random process that considers the current situation, as shown in Figure 1, rather than past events. The Markov property describes this form of memory loss [27].

There are several applications of the Markov chain in real models [28]. The Markov chain possesses the Markov property and is a random procedure that is discrete in time. However, some researchers refer to ongoing operations across time as Markov chains. In a time-discrete random operation, a system is in a predetermined state at each step and randomly changes conditions at each phase [29]. Stages are often related to time; however, they may also relate to a physical distance or different discrete numbers. The system's conditional likelihood distribution in the subsequent stage relies solely on the system's present state, not its previous circumstances, following Markov's property [30]. Because the system is unpredictable, it's nearly difficult to forecast the state of a Markov chain at a given point. On the other hand, the system's statistical characteristics are foreseeable in the future. These statistical features are crucial in many applications [31].

Transfer probabilities are the probabilities given to these state changes, and transmission systems are the likelihood ascribed to them. A collection of states and the possibility of transitions characterize a Markov chain, respectively. Due to the contract, we presume that the next stage is often available, and the process never ends.

The position varies with each step with an equal chance of +1 or -1 in the random walk, which is one of Markov's most prominent chains. There are two ways to go on from each place: one moves back to the integer that came before it (-1), and the other moves on to the integer that comes after it (+1). The current situation solely determines any transfer's likelihood. For instance, the possibility of migrating from 5 to 6 equals the probability of transferring from 5 to 4, equivalent to 0.5. The previous circumstance has no bearing on these possibilities (4 or 6). A series of random variables, such as X_n, that exhibit Markov characteristics, i.e.

In accordance with Eq.(1), the state space is a countable set of all possible values for X_i [32]. There is also the following definition.

$$\begin{aligned}&\Pr\left(X_{n+1}=x\,|X_1=x_1,X_2=x_2,\ldots,X_n=x_n\right)\\&=\Pr\left(X_{n+1}=x\,|X_n=x_n\right)\end{aligned}\quad(1)$$

If a repairable system cannot be considered to be fully repaired instantly, or, to put it another way, with a brief and insignificant length relative to the system's operating time, methods like Markov continuous chain are used for evaluation. The time or state space is constantly and discontinuously partitioned in Markov's technique for describing random manners. Random processes are continuous or discontinuous random changes. The Markov approach, in reality, demands that the system imitate memory deficiency. Consequently, the system's future state is unrelated to its previous forms and is solely dependent on its last component.

Since the researcher has limited time to analyze the data and must rely on sample data that reflect the population, access to the population is essential for sampling.

Necessary sampling procedures are shown in Figure 2. The first stage is to figure out where you are going. The sapling framework should then be used to pick a sample instance of the whole community. In relational theory, the definition of the sample is the selection of a subset from the sample framework or the total population and using the resulting sample instead of the dataset or generalization. The sampling procedure has a significant impact on this.

Figure. 1. Markov chain states

To avoid inaccurate or misaligned sampling, a sample should be reasonably sized. Overall, the sample size is important in this case, concerning the population's complexity, not the proportion of the sampled research population. Researchers' main goal is to study data; the more samples used, the less likely it is that the data would be distorted [33]. The sources used by the researcher state that when samples are more significant than a specific size, a yield loss occurs that has to be balanced [34]. A large sample size minimizes sampling error and slows the process.

Sampling the network and inferring its structural and behavioral characteristics from the sample data set is a way to reduce the network's complexity [35]. In the processing of network description and investigation, sampling approaches play a significant role. Sampling can analyze a tiny portion of a network while keeping the major network's properties [36]. The sample size is calculated using a variety of statistical procedures. Several formulae for computing the sampling size of categorized data are among the numerous techniques for determining the sample size. A large sample, nevertheless, cannot ensure accuracy.

$n = p\,(10^2 - p)\,z2 / E2$

n = Means the necessary sample size
p = Indicates the percentage of a circumstance or condition
e = Denotes the portion of the required maximum error
z = Shows the value that is determined by the level of confidence needed

1.2 The fundamentals of computer network sampling

Pairs of Γ = (ν, ε) are used to define the leading network, ν stands for nodes, and ε indicates groups of nodes that are linked to one another. The result of the G network is an incomplete complex network when the leading network is sampled. Navigation nodes are with their surrounding nodes. A sampling subnet from the top network is shown in Figure 3. The top-level network is represented by the topper net, while the lower-level network is an example of a randomly generated subnet with the insufficient specification.

1.3 Random walk on network

Equation (2) gives the following definition for the transfer matrix $p=[p_{ij}]$ of the random walk. Since it is a Markov chain, $d(i)$ is the angle of node i and $p=p_{ij}$ is the likelihood that the scroll will go from node i to node j. Similarly, p represents the likelihood of seeing each node at each stage. In line with the Markov chain's main theorem, there exists a single probability distribution $\pi=(\pi_1,\pi_2,\ldots,\pi_n)$, where π_i is equal to [37].

$$pij \begin{cases} 1/d\,(i) \in \varepsilon \\ 0 otherwise \end{cases} \qquad (2)$$

π_i in Eq.(3) is proportional to the angle, $i\ k\ (i)$, because ε is a constant, where π_i is the likelihood of encountering node i in a random walk with a given distribution. So, it is easy to sample the nodes with the most significant angle. In other words, the average angle of the subset must be greater than or equal to the angle of the primary grid. The simulation we ran and this outcome is identical [37].

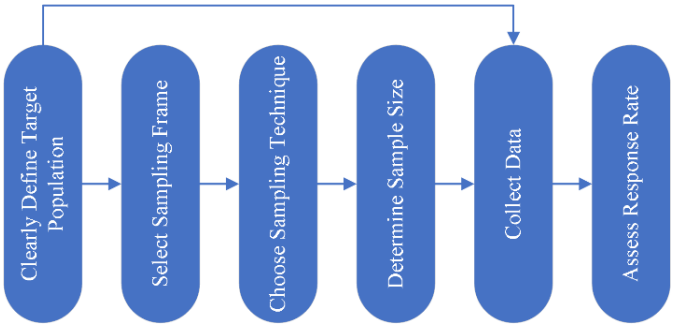

Figure. 2. Crucial sampling procedures [5]

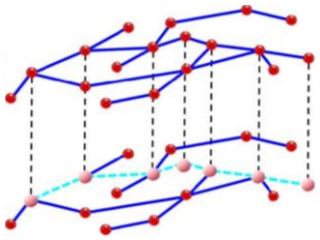

Figure. 3. The leading network is the top network, while the sample network is the bottom network. The return route is shown by big balls and dashed lines [10].

$$\pi_i = \frac{k\,(i)}{2|\varepsilon|} \qquad (3)$$

1.4 Page rank algorithm

An algorithm known as a page rank rate search engines' results which are websites that are specifically targeted for searches higher than other websites. The English name for this idea is derived from Larry Page, who created the algorithm and is one of the original Google founders [38]. With the help of this algorithm, anyone looking for a specific phrase might first view websites that are more popular and more relevant to them.

Page rank is calculated by examining the incoming connections to a website. Although this value is close to 10, it still implies many inbound links and outgoing connections to this website. Only a few websites worldwide rank 10, including Google.com and Wikipedia.com [39]. The page rank values of the websites it refers to, as well as the number of connections those pages have left behind, define the page rank value of a web page. According to Eq.(4), the purpose of this method is to give greater weight to sites with more referrals. The PageRank algorithm has the advantage of taking the referral page into account and the number of page references when determining relevance [40]. A page's page rank is determined using the Eq.(4).

$$PageRank\,(u) = (1-d) + d * \sum v \in B\,(u)\ PR\,(v)/N \qquad (4)$$

If a page is critical, it is placed at the top of the structured sorting model. The internal links of the web site should be raised if the pages have many outward links to maintain the same page rank. Increased internal linking has no effect on page rank if any site's pages lack an external connection. So, take care to avoid making your website inaccessible. Other businesses than Google.com employ this technique as well. Another example is Alexa.com, which uses an algorithm to

rank websites according to the actual number of visitors. The higher your website ranks, based on a rating of numbers between one and several million, the closer this number is to one. Bing.com has become a crucial component of modern worldwide search [41].

Because of this, web designers often concentrate on prominent search engines to build a website that is more suited to be ranked for that specific engine. Naturally, it will probably drop in some other engines' ranks. Some designers could even ignore the rating of general engines instead of focusing on a niche, regional, and specialized engines. Your website should be optimized to get the highest possible score in the search engine you are targeting before making any changes to its settings. With the above explanation, it becomes vital to be more cautious about the services offered to rank your website high in hundreds of thousands of search engines for a reasonable price [42].

1.5 Different random walk techniques

The node core is randomly chosen from the top network in the Classical Random Walk approach. In the following step, node i neighbors are randomly chosen to replace them with node j. Choosing the initial node and selecting the most suitable route to follow are the two main variables that influence scrolling. Two modified random walking algorithms may be considered to research these two elements [10]:

a) Selecting the Seed Node

The seed node is selected to match the node's degree. The walker proceeds in the same way as a regular random walk. This increases the likelihood of a high-degree node being selected in the sampling's first stage. Therefore, this technique emphasizes the importance of a high-degree node.

b) No-Retracting

This method determines the seed node and the subsequent nodes, except the most recent node visited, from among those nearby. This makes it obvious how "going back" impacts the effectiveness of the sampling.

1.6 Distribution of degrees

The initial degree of a node is determined in the field of networks and diagrams by the number of links it has to other nodes, as well as the distribution of the expected degrees that these connections have throughout the network. The number of connections between nodes makes up a node's degree in a network, which is frequently used interchangeably with the word "node" [43]. Edges in a directional graph are links that lead from one node to another. As a result, each node will have two degrees: an input degree that indicates how many input edges are present on the node and an output degree that shows how many output edges are present on the node. The presented formula in Eq.(5) can calculate the degree of distribution [44].

$$P\left(k\right) = {n_k}/{n} \qquad (5)$$

If a portion of the nodes have degree k and have the degree of distribution *p*, in Eq.(5), *P(k)* will be the degree of distribution of a network with *n* nodes and n_k nodes with degree *k*. Degree distribution enables the completion of two activities.

1.It demonstrates the distribution of connections among network nodes or that each node has many connections with its neighbors.

2. It also demonstrates how the probability is dispersed throughout the network's other nodes.

1.7 Average degree

The average number of edges per node is the definition of the average degree in a diagram [45]. Eq.(6) is used to compute it, which is a relatively straightforward process [46].

$$ave\_deg = \frac{TotalEdges}{TotalNodes} = \frac{m}{n} \qquad (6)$$

1.8 Clustering coefficient

The number of nodes in a graph tends to group is measured by the clustering coefficient in graph theory. Evidence reveals that nodes often form groups of the same kind with rather intense communication in real-world networks, particularly social networks. This concept is more likely than the average chance of connections forming randomly between two nodes [47].

Consideration of the Clustering coefficient formulae, which may be seen as a starting point for a critical communication analysis, is one of the most acceptable methods to discover mass communication inside a group [48]. A node v_i with degree K_i*'s* clustering coefficient is calculated in Eq.(7).

$$C_{V_i} = 2 * \frac{L_i}{K_i\left(K_i - 1\right)} \qquad (7)$$

L_i represents the total number of connections between node v_i*'s* neighbors, except the node itself. Cv_i*'s* value is consistently between [0, 1]. If $Cv_i = 0$, then none of node v_i*'s* neighbors are linked to any other neighbors, and if $Cv_i = 1$, then all of node v_i*'s* neighbors are connected.

2. Previous Work

Two different random walks, picking seed nodes and no-retracing, have been suggested by Xiea et al. [10]. These strategies focus on the implications of the seed node selection and route overlap instead of standard random walks. Based on the proposed results, efficiency has increased using this approach.

Tong et al. [49] have investigated complex networks, including social networks, which have gained popularity recently. Due to the results more comprehensive network investigations are required about processing power, storage space, and energy usage because of the scale and complexity of these networks. This research has developed green sampling algorithms to lower energy usage and increase efficiency.

According to Fang et al. [50], a graph-based method for estimating picture saliency takes into account explicit visual

cues and latent signal correlations. The authors repeatedly performed a human-supervised random walking algorithm on graphs to improve the accuracy of the findings.

Rezvanian et al. [51] used the shortest route theory to sample social networks. The suggested sampling strategy starts by identifying the shortest routes between several node pairs. Then the edges in these routes are then ranked by how frequently they occurred. Finally, a subgraph with high-rated edges is calculated from the sampled network.

3. Proposed Method

This section introduces our proposed method. Figure 4 depicts the stages involved in running the algorithm. After receiving the input data, three algorithms are applied, including random walk, picking seed node, and the no-retrace. Then, these algorithms are contrasted one by one.

Since it is possible to infer the network's characteristics based on the network properties that the user is interested in, the components of the network may be evaluated with little testing and study among typical methodologies and sampling. However, the researcher must be aware of the ideal sample volume and the optimum strategy for accelerating the sampling process via repeated testing. A technique that combines page ranking and random navigation algorithms is proposed to make graphs easier to understand and shorten trip routes.

The procedures for running these algorithms are shown in Figure 5. The procedure is the same as the prior algorithm, except that data is processed first by *Google's* PageRank.

4. Experimental Results

In this section, the approaches that were used are evaluated along with the suggested results. This research used data from the American airline network in 1997 (USAir97) [52]. The dataset's 332 nodes include information like each node's connection to terminals (degree distribution).

We examined numerous factors randomly removed from the sample network to get the average across the whole network [53]. The complete version of the USAir97 with 332 nodes was used to generate a graph in Figure 6. Figure 7 displays the database's clustering coefficient. Additionally, Figure 8 illustrates that to an average level.

4.1 Application of the standard random walk

The traditional random walk is implemented using a class named Random Walk. This class takes a list of nodes and edges as input and returns a list of nodes that this function has visited. Before running this operation, a timer is started to record how long the scrolling takes. The time is then logged after scrolling. When the resultant nodes have been turned into grid edges and the graph has been redrawing, the timer is stopped. This results in a traversal time and diagram drawing time.

4.2 Classic random walk using with PageRank

The Google PageRank algorithm operates on the information received from the input. Moreover, its key nodes are identified. Its algorithm allows for adjusting the number of these nodes. In this experiment, the default selection is made up of five nodes. Finally, a random scrolling mechanism is used to scan the new data.

4.3 Seed node selection

The seed node selection technique sorts the nodes by their most crucial degree before scrolling them using the Random Walk class. Following navigation utilizing the ninety-core selection approach, the United States flights are shown in Figure 9, demonstrating how the model structure and node distance have changed.

4.4 Seed node selection with PageRank

Figure 10 also displays the graph after selecting the seed node and navigating the page rank combination. The graph demonstrates that nodes with a higher degree are more likely to be sampled.

4.5 No-retracing

This technique was implemented using the random_walksNR class, with modifications to its return function. The next node in this procedure is chosen randomly among a node's neighbors, except for the case that was visited in the previous phase. This approach makes it clear how backtracking affects sampling effectiveness. A graph produced using the no-retracing method is shown in Figure 11.

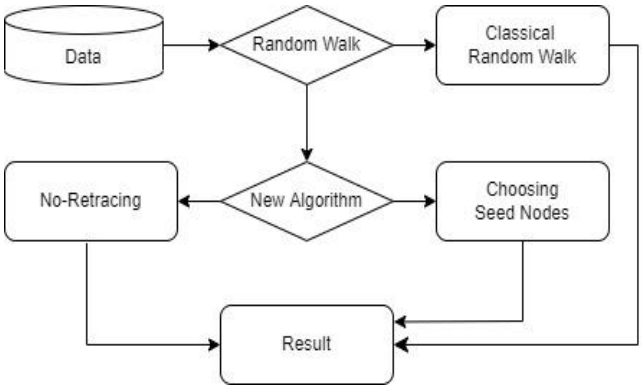

Figure. 4. Actions to do before to the page ranking algorithm

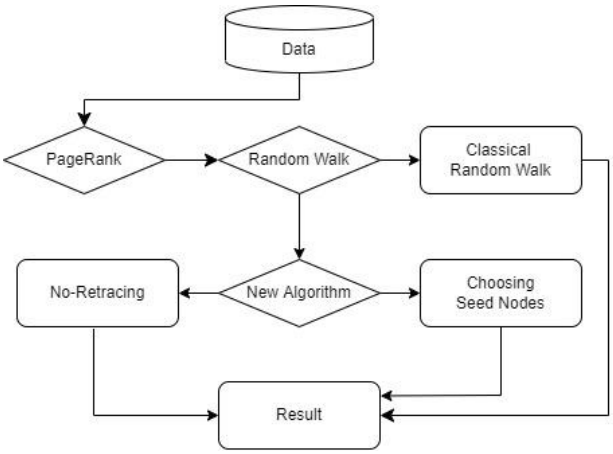

Figure. 5. Actions to do after the page ranking algorithm

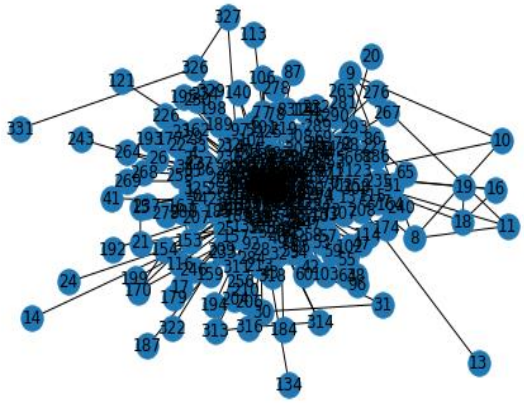

Figure. 6. 332 nodes were used to generate a graph from the USAir97

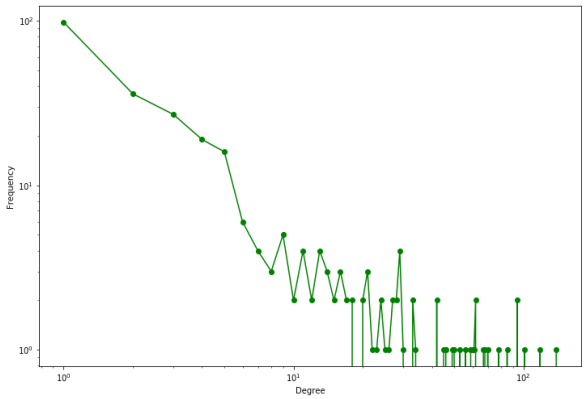

Figure. 7. Clustering coefficient in the 332 nodes USAir97

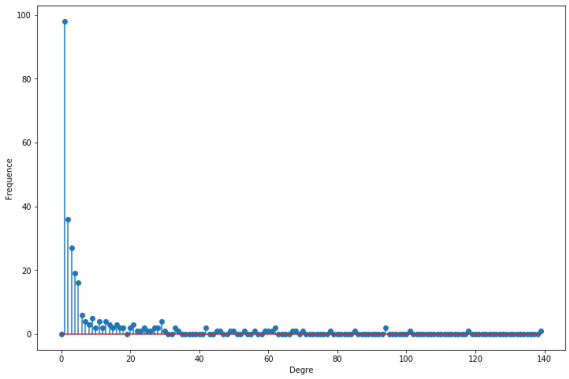

Figure. 8. USAir97 average degree with 332 nodes

4.6 Using No-retracing with PageRank

This method was developed by fusing a no-retracing technique with a PageRank algorithm. Except for the sample, which often includes nods of a greater degree, the resultant graph is comparable to the graph produced by the no-retracing method. A graph was created using the page rank algorithm, and no retracing is shown in Figure 12.

5. Conclusion

Benefits and downsides for each of the described sampling techniques, and the optimum method should be chosen based on the resources at hand. The time required to

Figure. 9. Choosing the seed node algorithm for creating the graph

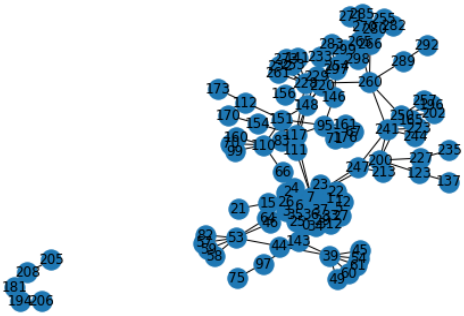

Figure. 10. Choosing the seed node with the page rank algorithms for creating the graph

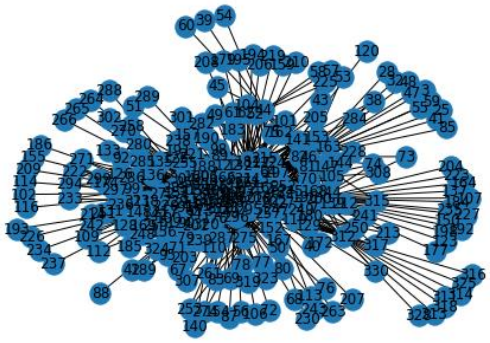

Figure. 11. Choosing the no-retracing algorithm for creating the graph

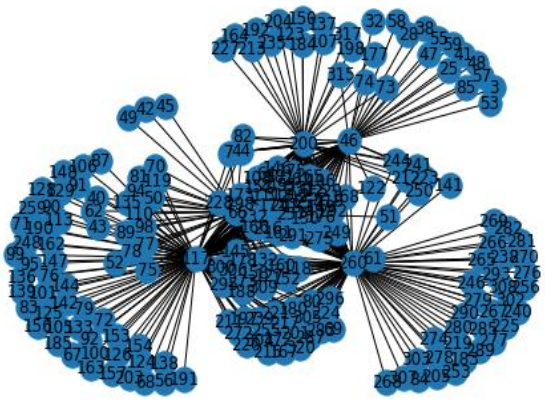

Figure. 12. Choosing the seed node with the no-retracing algorithms for creating the graph

execute the examined algorithms entirely is shown in Table 1. It is evident that among these algorithms, the PageRank algorithm with the choice of seed node selection has the quickest speed to finish the process, with a length of 10,362 milliseconds. The slowest method is the conventional random walk, which has a period of 22,901.

It turns out that sampling criteria go beyond just time. The accuracy standard is another essential requirement. The sample is more closely aligned with the leading network, the better the sampling accuracy. Accuracy and speed are often negatively linked. As a result, consideration must constantly be given while choosing the best algorithm for the circumstances and resources.

The mean degree and coefficient of the investigated algorithms are shown in Table 2. The classical random walk method has the most significant average degree of these algorithms, 11.081, demonstrating a significant association between the model's nodes. The no-retracing technique has the most effective average degree (0.587), indicating that this model's nodes strongly desire to form clusters and interact with nearby nodes. The algorithm's accuracy increases with the size of the criteria under consideration. The optimum method is thus one that executes more quickly and differs little from the original sample and the assessed criteria.

Table 1. The Algorithms Under Examination's Time in Milliseconds

Name of the Algorithm	***Duration (milliseconds)***
Classic random walk	**22,901**
Classic random walk + PageRank	15,421
Seed node	18,819
Seed node + PageRank	**10,362**
No-retracing	22,316
No-retracing + PageRank	13,987

Table 2. Comparing of Methods' Average Degrees and Clustering Coefficients

Name of the Algorithm	***Number of Edges and Nodes***	***Average Degree (AD)***	***Clustering Coefficient (CC)***
Classic method of random walk	1496, 270	**11.0810**	0.5520
Classic method of random walk + PageRank	1496, 270	8.9460	0.5340
Seed node selection	278, 139	4.0000	0.2920
Seed node selection + PageRank	130, 106	2.4520	0.0940
No-retracing	1590, 267	11.9100	**0.5870**
No-retracing + PageRank	450, 2019	4.1090	0.4570

Declarations

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors' contributions

AB: Study design, interpretation of the results, statistical analysis, drafting the manuscript and supervision; AK: Study design, design, interpretation of the results, drafting the manuscript, revision of the manuscript; AF: Drafting the manuscript, interpretation of the results, revision of the manuscript;

Conflict of interest

The authors declare that there is no conflict of interest.

References

[1] D. Lusinchi, "“President” Landon and the 1936 Literary Digest Poll: were automobile and telephone owners to blame?," *Social Science History*, vol. 36, no. 1, pp. 23-54, 2012.

[2] S. L. Lohr and J. M. Brick, "Roosevelt predicted to win: Revisiting the 1936 Literary Digest poll," *Statistics, Politics and Policy*, vol. 8, no. 1, pp. 65-84, 2017.

[3] J. ScottArmstrong, "Why the 1936 literary digest poll failed: Peverill Squire, Public Opinion Quarterly 52 (1988) 125-133," *International Journal of Forecasting*, vol. 5, no. 2, pp. 295-295, 1989.

[4] I. Etikan and K. Bala, "Sampling and sampling methods," *Biometrics & Biostatistics International Journal*, vol. 5, no. 6, p. 00149, 2017.

[5] H. Taherdoost, "Sampling methods in research methodology; how to choose a sampling technique for research," *International Journal of Academic Research in Management (IJARM)*, vol. 5, no. 2, pp. 18-25, 2016.

[6] S. Zeadally, G. Martinez, and H.-C. Chao, "Securing cyberspace in the 21st century," *Computer*, vol. 46, no. 04, pp. 22-23, 2013.

[7] Z. S. Jalali, A. Rezvanian, and M. R. Meybodi, "Social network sampling using spanning trees," *International Journal of Modern Physics C*, vol. 27, no. 05, p. 1650052, 2016.

[8] S. Fottovat, H. Izadkhah, and J. Hajipour, "Community Detection in Social Networks Considering the Depth of Relationships," in *2022 8th International Conference on Web Research (ICWR)*, IEEE, 2022, pp. 24-28.

[9] O. Frank, "Network Sampling," in *International Encyclopedia of Statistical Science*, M. Lovric Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 941-942.

[10] Y. Xie, S. Chang, Z. Zhang, M. Zhang, and L. Yang, "Efficient sampling of complex network with modified random walk strategies," *Physica A: Statistical Mechanics and its Applications*, vol. 492, pp. 57-64, 2018.

[11] M. Hajarian, A. Bastanfard, J. Mohammadzadeh, and M. Khalilian, "SNEFL: Social network explicit fuzzy like dataset and its application for Incel detection," *Multimedia tools and applications*, vol. 78, no. 23, pp. 33457-33486, 2019.

[12] L. L. Peterson and B. S. Davie, *Computer networks: a systems approach*. Elsevier, 2007.

[13] A. R. Rohani and A. Bastanfard, "Algorithm for persian text sentiment analysis in correspondences on an e-learning social website," *Journal of Research in Science, Engineering and Technology*, vol. 4, no. 01, pp. 11-15, 2016.

[14] M. Hajarian, A. Bastanfard, J. Mohammadzadeh, and M. Khalilian, "Introducing fuzzy like in social networks and its effects on advertising profits and human behavior," *Computers in Human Behavior*, vol. 77, pp. 282-293, 2017.

[15] R. Dolan, J. Conduit, C. Frethey-Bentham, J. Fahy, and S. Goodman, "Social media engagement behavior: A framework for engaging customers through social media content," *European Journal of Marketing*, vol. 53, no. 10, pp. 2213-2243, 2019.

[16] M. Hajarian, A. Bastanfard, J. Mohammadzadeh, and M. Khalilian, "A personalized gamification method for increasing user engagement in social networks," *Social Network Analysis and Mining*, vol. 9, no. 1, pp. 1-14, 2019.

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

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

[19] P. Erdos and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, no. 1, pp. 17-60, 1960.

[20] E. N. Gilbert, "Random graphs," *The Annals of Mathematical Statistics*, vol. 30, no. 4, pp. 1141-1144, 1959.

[21] J. Hughes, *SAGE Internet Research Methods: Policy, Professionalism and Change* (SAGE Internet Research Methods). London: Sage Publications Ltd, 2012.

[22] S. K. Thompson, "Adaptive web sampling," *Biometrics*, vol. 62, no. 4, pp. 1224-1234, 2006.

[23] M. P. Stumpf and C. Wiuf, "Sampling properties of random graphs: the degree distribution," *Physical Review E*, vol. 72, no. 3, p. 036118, 2005.

[24] A. K. Moghadam and A. Bastanfard, "Improving Random Walk Sampling, Inspired by Two Methods of Choosing Seed Node And No-Retracing With Combination of them with Page Rank Algorithm," in *2022 8th International Conference on Web Research (ICWR)*, IEEE, 2022, pp. 195-202.

[25] K. Pearson, "The problem of the random walk," *Nature*, vol. 72, no. 1865, pp. 294-294, 1905.

[26] C. J. Geyer, "Practical markov chain monte carlo," *Statistical science*, vol. 7, no. 4, pp. 473-483, 1992.

[27] G. L. Jones and Q. Qin, "Markov chain Monte Carlo in practice," *Annual Review of Statistics and Its Application*, vol. 9, pp. 557-578, 2022.

[28] D. S. Myers, L. Wallin, and P. Wikström, "An introduction to Markov chains and their applications within finance," *MVE220 Financial Risk: Reading Project*, 2017.

[29] E. Behrends, *Introduction to Markov chains*. Springer, 2000.

[30] D. W. Stroock, *An introduction to Markov processes*. Springer Science & Business Media, 2013.

[31] K. C. Chan, C. Lenard, and T. Mills, "An introduction to Markov chains," in *Proceedings of the 49th Annual Conference of Mathematical Association of Victoria*, J. Cheeseman, Ed., 2012: Mathematical Association of Victoria, in It's my maths: personalised mathematics learning, pp. 40-47.

[32] S. Ghahramani, *Fundamentals of probability: with stochastic processes*. Chapman and Hall/CRC, 2018.

[33] G. Sharma, "Pros and cons of different sampling techniques," *International journal of applied research*, vol. 3, no. 7, pp. 749-752, 2017.

[34] A. M. Adam, "Sample size determination in survey research," *Journal of Scientific Research and Reports*, vol. 26, no. 5, pp. 90-97, 2020.

[35] C. Phillips, "Sample size and power: What is enough?," in *Seminars in Orthodontics*, 2002, vol. 8, no. 2: Elsevier, pp. 67-76.

[36] M. Papagelis, G. Das, and N. Koudas, "Sampling online social networks," *IEEE Transactions on knowledge and data engineering*, vol. 25, no. 3, pp. 662-676, 2011.

[37] L. Rodero-Merino, A. F. Anta, L. López, and V. Cholvi, "Performance of random walks in one-hop replication networks," *Computer Networks*, vol. 54, no. 5, pp. 781-796, 2010.

[38] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford InfoLab, 1999.

[39] P. Berkhin, "A survey on PageRank computing," *Internet mathematics*, vol. 2, no. 1, pp. 73-120, 2005.

[40] M. Perc, "The Matthew effect in empirical data," *Journal of The Royal Society Interface*, vol. 11, no. 98, p. 20140378, 2014.

[41] B. J. Jansen and A. Spink, "How are we searching the World Wide Web? A comparison of nine search engine transaction logs," *Information processing & management*, vol. 42, no. 1, pp. 248-263, 2006.

[42] A. Spink, *Web search engine research*. Emerald Group Publishing, 2012.

[43] X. Wang, F. Ma, and B. Yao, "Arbitrary degree distribution networks with perturbations," *AIP Advances*, vol. 11, no. 2, p. 025301, 2021.

[44] S. N. Dorogovtsev and J. F. Mendes, "Evolution of networks," *Advances in physics*, vol. 51, no. 4, pp. 1079-1187, 2002.

[45] A. Yang, X. Huang, X. Cai, X. Zhu, and L. Lu, "ILSR rumor spreading model with degree in complex network," *Physica A: Statistical Mechanics and Its Applications*, vol. 531, p. 121807, 2019.

[46] G. Ayyappan, C. Nalini, and A. Kumaravel, "A study on SNA: measure average degree and average weighted degree of knowledge diffusion in Gephi," *Indian Journal of Computer Science and Engineering*, vol. 7, no. 6, pp. 230-237, 2016.

[47] X. Pan, G. Xu, B. Wang, and T. Zhang, "A novel community detection algorithm based on local similarity of clustering coefficient in social networks," *IEEE Access*, vol. 7, pp. 121586-121598, 2019.

[48] A. Said, R. A. Abbasi, O. Maqbool, A. Daud, and N. R. Aljohani, "CC-GA: A clustering coefficient based genetic algorithm for detecting communities in social networks," *Applied Soft Computing*, vol. 63, pp. 59-70, 2018.

[49] C. Tong, Y. Lian, J. Niu, Z. Xie, and Y. Zhang, "A novel green algorithm for sampling complex networks," *Journal of Network and Computer Applications*, vol. 59, pp. 55-62, 2016.

[50] M. Fang, J. Yin, and X. Zhu, "Active exploration for large graphs," *Data mining and knowledge discovery*, vol. 30, no. 3, pp. 511-549, 2016.

[51] A. Rezvanian and M. R. Meybodi, "Sampling social networks using shortest paths," *Physica A: Statistical Mechanics and its Applications*, vol. 424, pp. 254-268, 2015.

[52] V. Batagelj and A. Mrvar. *Pajek datasets*. [Online]. Available: http://vlado.fmf.uni-lj.si/pub/networks/data/

[53] R. Rossi and N. Ahmed, "The network data repository with interactive graph analytics and visualization," in *Twenty-ninth AAAI conference on artificial intelligence*, USA, AAAI Press, 2015, vol. 29, no. 1, pp. 4292-4293.

Azam Bastanfard is an Assistant Professor at the Islamic Azad University, Karaj branch. She received her Master's degree and Ph.D. from the Tokyo Institute of Technology. She had a postdoctoral position at the University of Geneva. Her field of study is Multimedia Processing, Artificial Intelligence, Gamification, and Social Networks.

Ali Kheradbeygi Moghadam received his Bachelor's degree in computer software engineering in 2017 at the Islamic Azad University, Takestan branch. He received his Master's degree in Computer Software Engineering in 2020 from the Islamic Azad University, Karaj Branch. His research areas are Social Networks, Software Development and Image Processing.

Ali Fallahi RahmatAbadi received his Bachelor's degree in computer software engineering in 2014. He began researching Recommender Systems in 2016 and received a Master's degree in Computer Software Engineering in 2018. Currently, he is a Ph.D. candidate in Computer Software Engineering. His research areas are Recommender Systems and Machine Learning.