Journal of Theoretical and Applied Information Technology
31st August 2017. Vol.95. No.16
© 2005 - Ongoing JATIT & LLS

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FAST SUMMARIZATION OF LARGE-SCALE SOCIAL NETWORK USING GRAPH PRUNING BASED ON K-CORE PROPERTY

[1,2] ANDRY ALAMSYAH, [1] YOGA PRIYANA, [1] BUDI RAHARDJO, [1] KUSPRIYANTO

[1] School of Electrical Engineering and Informatics, Bandung Institute of Technology, Indonesia

[2] School of Economic and Business, Telkom University, Indonesia

E-mail: andrya@telkomuniversity.ac.id

ABSTRACT

Graph based modelling is common in many implementation areas involving combinatorics relationship such as in social network. The data explosion produced from user generated content in online social network services trigger the emergence of large-scale social network. Having large graph at our disposal gives us many opportunity but at the same time increase the complexity problem, especially in several graph metric computations and also at graph visualization. A fast summarization methods is needed to reduce the graph size into the only most important pattern. This summarize sub-graph should represent the property or at least converge to the value of the original graph property. Social Network is characterized by scale free degree distributions, which have fat-head less important nodes that can be removed. Graph Pruning method is introduced to remove less important nodes in certain graph context, thus reduce the complexity of large-scale social network while still retain the original graph properties. The method is based on k-core graph properties. The paper show how is the effect of graph pruning to the several most used social network properties.

Keywords: *Social Network Analysis, Graph Pruning, Graph Theory, K-Core, Graph Sampling*

1 INTRODUCTION

The internet generates large volume of data, where later those data can be beneficial in constructing human behavioral model. Some real-world application quantifies internet-based human-social behavior for usage such as in business marketing [1], politics [2], knowledge management [3], dissemination information [4] and other areas. Today, many researches and applications prefer using internet-based behavioral measurement over previous approach using sampling or questionnaire, because the latter is cheaper and faster [5].

Most of internet data is in unstructured form. This can be a challenge when creating model using established method such as data mining. In general, there are two main methods deal with unstructured data; they are *Text Mining* (TM) [6] and *Social Network Analysis* (SNA) [7]. TM process text data into meaningful content. SNA process relationship data into structural network, where pattern of relationship is valuable to the model creation. SNA is considered faster methodology by capture only the relationship data. SNA borrows metrics and problem formulation from *graph theory* [8].

The large volume data mostly generate large-scale social network. Large graph as a representation of large-scale social network. Since most of graph metric is not built with scalability in mind, this can pose increasing complexity when data explode such as user generated data in internet. The problem includes some metric might not be able to be computed at all and graph visualization is hard to understand. For this reason, there are many researches concern to simplify the complexity. We will see the details of the researches effort in chapter 1.1

Developing algorithm that are accurate, scalable and efficient for social network is not easy [9]. A good approximation may be used to be able to process large-scale social network. Relaxing the requirement of metric computation is also could help to solve the problem. The approximation result is accepted if it reflects the original metric computation.

The distinct feature of the social network is their *scale free distributions*. It distances themselves from random network where we expect most nodes have around the same number of connections around an average, this form a uniform probability distributions (fig.1a). The *preferential attachment* mechanism [10] explain that new nodes who joined to the social network have higher probability connected to the high degree nodes in the network. The accumulative advantages of the highest degree nodes mean that the degree distribution will form an approximate *scale free distributions* (fig. 1b).

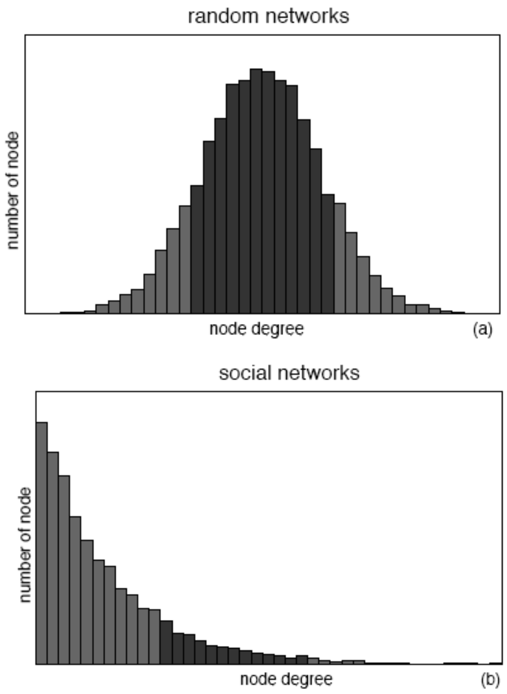


Figure 1. (a) random networks degree distributions. (b) social network degree distributions

1.1 Related Works

There are some approaches to reduce complexity of processing large graph depending on the contextual utilization, for example to measure a NP-hard metric on large graph, we can build new metric with an approach not to include all the nodes and / or edges in the computation. Another way is to reduce the graph size before the actual processing. The most common way to reduce graph size is through Graph Sampling.

Graph Sampling is process of finding representatives subset of the original graph. We call the representatives subset as sample graph [11] [12]. This process requires that the sample graph should preserve the original graph properties. Several important methods for graph sampling process are based on edge random sampling [13] [14], node random sampling [13] [14], random walk sampling [13] [15], and combinations between those methods [16]. Each of the methods works for specific graph properties, for example to preserve graph connectivity our choice will be random walk sampling, while to preserve degree distributions we can use node and / or edge sampling.

1.2 Contributions

Our contribution based on the research question on how to build fast reduction methods of graph size, while still maintain the graph property accuracy. Our proposed methods require social network characteristics *scale free distributions*. The approach is by pruning less important node and / or edge in the graph. The cut-off filtering fat-head part of *scale free distributions* truncate graph size in very fast fashion compared with other methods such as deletion node and / or edge iteratively until reach certain size [13], contracting adjacent nodes [14], graph induction [13] and traversal based sampling [15].

Graph pruning is based on *k-core* graph property. *k-core* decomposition used as basis for filtering out degree node less than *k*, thus we can focus on the most important structures of the social network [17] for further process such as metric computations or graph visualization. Graph sampling is still an exhaustive process when the graph is too large. It needs certain pre-processing work. Graph pruning can efficiently fulfill this requirement. Efficient in term of fast processing and produce accurate representation of large-scale social network for further graph analysis, including graph sampling.

One conditions where our solution required is when we want to compute path-length-based graph metric such as *betweenness centrality, average path length,* or *diameter* of large-scale graph. Their computation complexity reach $O(n^3)$, where n is number of nodes [18]. By using our methodology, we can quickly prune the large number of nodes, leaving the core node intact and easing the computation complexity.

2 THEORITICAL FOUNDATIONS

2.1 Social Network Analysis

Social Network Analysis is graph representation of relationship between actors. The actors represented as nodes, while the relationship between actors are represented as edges [7]. A graph $G(N.E)$ is the formal representation of a social network, N is the set of nodes and $E \subseteq \{(u, v)|u \in N, v \in N\}$ is the set of edges, where (u, v) is an unordered pair of nodes. We denote $n = |N|$ and $m = |E|$. The neighbors of node v is defined as set

NG(v) = {u| (u, v) ∈ E, u ∈ N}. The degree of a node *v* is defined as $d_G(v) = |NG(v)|$.

In Fig. 2, a graph illustration *G(N,E)* is shown with set of nodes *N = {1,2,3,4,5,6,7}* and set of edges *E ={(1,2).(1,5).(2,5),(3,4),(5,7)}*. We have *7* nodes and *5* edges. We take an example of the degree of node *5* is $d_G(5) = 3$.

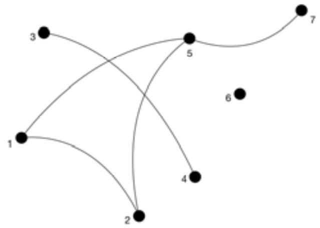


Figure 2. A Graph Illustration Of 7 Nodes And 5 Edges

Having graph theory as a base of social network formulations, we have the advantage of constructing several graph metrics to quantify social network. Some the most used properties in social network are as follows [19]:

1. *Average Degree* of a network *G* is measured by compared number of edges *E* to number of nodes *N*. We denote Average Degree *AvgDeg = |E|/|N|*
2. *Density* is measured by comparing actual number of edges *E* in the network *G* to maximum possible number of edges. We denote *Den = |E|/(|N|*(|N|-1))*
3. *Average Path Length* is calculated by finding the shortest path between all-pairs of nodes, adding them up and then dividing by the total number of pairs.
4. *Degree Distributions* measures network characteristic by the distributions of node degree in the network. Some compare node degree and its appearance frequency and some compare node degree and their fraction node to the overall number of nodes. Both measures the same thing.
5. *Average Clustering Coefficient* measures the total degree to which nodes are tend to cluster together in network compared to total number of nodes
6. *Diameter* measures the longest of shortest path between any pair of nodes in network
7. *Centrality* measure the most important nodes based on the certain context such as the most connected node *(degree centrality),* node that has the most number of shortest path going through them *(betweenness centrality)*, and node that has average shortest distance to any other nodes *(closeness centrality).*
8. *Modularity* measures fraction of the edges that fall within the given group minus the expected of such fraction of edges were distributed at random. The bigger *Modularity* value means the boundary between groups in the network are more distinct.
9. *Connected Component* show how many network component in which any two nodes are connected to each other by the existence of paths.

We classify those network metrics above into three categories: single value measurement, rank measurement, and distributions measurement. *Average Degree, Density, Average Path Length, Diameter, Modularity, Connected Component* are a single value measurement, while *Centrality* is rank measurement, and at last the *Degree Distributions* is a distributions measurement. The high complexity of node relationship and graph topology cannot always be captured by a single value measurement, that is why in this paper, we measure the effect of pruning process to all those three categories.

2.2 Graph K-Core Property

Definition: A subgraph *G'(C, E|C)* induced by the set *C* ⊆ *N* is a *k-core* if and only if the degree of every node *v* ∈ *C* induced in *G'* is greater or equal than *k*. This can be read as $\forall v \in C$: $d_{G'}(v) \geq k$, and *G'* is the maximum subgraph with this property [16].

The *k-core* is the subgraph obtained from the original graph by the recursive removal of all nodes of degree less than or equal than *k*. The node coreness k_c of a given node in *c* is the maximum *k* such that this node is present in *k-core* graph, but removed from the *(k+1)-core* graph.

k-core provide a mean to identify internal network cores and a recursive process of network decomposition from the least important to the more important sub network. We regard this measure as an indicator of node centrality, since it measures how deep within the network a node is located. Another good feature of *k-core* property is that it is easily to compute. According to [20] *k-core* can be implemented with as low complexity as *O(l)*, where *l* is number of code lines.

2.3 Graph Pruning

Based on *k-core* definition, we derived the following proposition of Graph Pruning methodology.

Proposition: Let set *A* = {a_1, a_2, a_3, …, a_n} is a topological graph property of $G(N_G,E_G)$. Graph

$G'(N'_G, E'_G)$ is a product of pruning process. Set of topological graph property *$A'=\{a'_1, a'_2, a'_3,\ldots, a'_n\}$* of graph *G'*. Given an error *ε*, we can prune a set of nodes *$\{n_1, n_2, n_3,..., n_i\}$* and / or a set of edges *$\{e_1, e_2, e_3, \ldots, e_j\}$* that are the least important nodes and / or edges in *G* with the condition that fulfil *$A\text{-}A' < \varepsilon$*

Proof: Since we are looking for graph *G'*, a smaller representatives of graph *G*. We need to make sure that *G'* contain of network core which topological graph property value *A'* dominate the proportion of topological real graph property value *A*. This is guarantee by social network *Barabasi-Albert* model [10], Thus given an error *ε* then we guarantee that *$A\text{-}A' < \varepsilon$*

We note several global issues regarding the pruning process: (a) how to define the least important nodes and / or edges in the network, (b) what are the requirements of pruning process, (c) decide what criterions to stop the pruning process, some of the candidate of stopping pruning process are graph connectivity or the number of connected component (see 2.1), directly prune the fat-head area of scale free degree distributions, and checking metric comparison iteratively on each prune step.

For the case of *scale free distributions* of social network, we can easily prune the fat-head part of distributions. A given error *ε* become stopping criteria of the pruning process. In this paper, we focus on node degree context to define the node importance in the network. We can always use other property to define node importance such as more complex centrality measurement. Following the facts above, *scale free distributions* of social networks and stopping criteria *ε* are becoming the requirement of graph pruning process.

2.4 Analytical Approach Through Case Study

To have better illustration on how graph pruning performs, we use a case study to see the implementation of graph pruning through an analytical approach. The case study is the combination graph pruning implementation as pre-process step to reduce graph size and general graph sampling random walk as the main process to get the sampling graph accurately.

In sub chapter 1.1, we have shown that graph sampling can be used to summarize large-scale social network. Two prominent graph sampling based on random walk methods are *Metropolis Hastings Random Walk* (MHRW) [16] and *Forest Fire* (FF) [13].

MHRW built on the purpose to get unbiased graph representation. This can be achieved by introducing a checking mechanism on each iteration whether to accept or refuse next walker node destination. The acceptance or refusal state based on the graph property value of next node converge to the original graph or not. If the value is converging to the original value then we accept the walker to move to the next node, but otherwise we refuse the walker movement. MHRW mechanism guarantee the accuracy of the graph sample.

FF sampling mechanism work in parallel fashion, where instead the walker move to one node at a time, they can move to more than one node at a time. Comparing to traditional random walk sampling, the FF sampling run faster.

Random Walk Sampling Definition: Let $X = \{x_1, x_2, x_3, \ldots, x_n\}$ random walk node sequence on graph G. t is the time to collect X and construct subset graph G'. Let η(G) is a set of topological graph properties of G, then the objective is $\eta(G') \approx \eta(G)$.

Analysis:

In FF, t' is the time to collect X. Because of the parallel nature of random walk collection process, then we have $t' < t$.

In MHRW, for each step of random walk collection process, it needs to check whether η(G') converge to η(G). The time needed for this process is t", thus we have $t < t"$.

Let |g| is the size of graph G.

Graph Pruning pre-process step reduce graph G into graph G' with the size |g'|, thus $|g'| < |g|$. Random Walk process produces subset graph G" from graph G'. Since graph G" contain only the core part of graph G, thus we have $\eta(G") \approx \eta(G)$.

3 RESEARCH METHODOLOGY

3.1 Problem Definition

Given a large-scale social network *S*, our objective is finding *S'*, a summary of network *S*. A distance between *S'* and *S* network properties are measured to evaluate the effectiveness of summarization methods on each pruning percentage (pruning step) or each summary graph size.

To see the overall performance, we measure the summarization effect on three metric classification in 2.1 that is single value measurement, rank measurement, and distributions measurement.

We investigate how graph properties value evolve on each decreasing summary graph size represented by each increasing pruning percentage step. We expect there are some speed variation on how fast a summary graph properties distance itself from the original properties. Other than monotonic relationship between increase or decrease graph properties and pruning percentage, we may also have several properties which behave independently to size factor.

3.2 Experiment Design and Measurement

Our experimental design based on the idea of graph properties evolvement regarding graph reducing size. We inspect the effect of graph size on several measurements. First is to single value measurement, which are *Average Degree, Average Path Length, Density, Modularity, Average Clustering Coefficient,* and *Diameter*. Second is to distributions measurement, which is *Degree Distributions*. Third is to rank measurement, which is *Centrality* family.

We measure the difference between graph properties value in original graph and in summary graph. On single value measurement, we measure distance between original graph properties values and each summary graph property values. On distribution measurement, we evaluate similarity between original and summary distribution curve by using *Discrete Frechet Distance*. While for rank measurement, we measure rank consistency by using *Kendall Rank Correlation Coefficient*.

Frechet Distance (FD) measure the similarity between curves [21]. The basic ideas arise from the question how to compare two shapes, whether it is a point sets, polygons, images, triangular, meshes, etc. The comparison measures the maximum distance between any comparable points in both curves. FD formal definition is as follows: Given two continuous curve α and β, then FD $\Delta(\alpha,\beta)$ is the minimum over all re-parameterization $f:[0,1]\rightarrow\alpha$, $g:[0,1]\rightarrow\beta$ of maximum over all $t\in[0,1]$ of the distance in $f(t)$ and $g(t)$, where f and g are continuous non decreasing function defining the positions of each comparable node in both curves α and β at every instant.

Kendall Rank Coefficient Correlation (KRCC) is a statistic to measure correlation of given two rank-order measurement [22]. If we have pairs observations (x_i,y_i) and (x_j,y_j), where $i\neq j$. Concordant if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$. Discordant if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$. The formulation is KRCC τ = *(number of concordant pairs) – (number of discordant pairs)* divide with $n(n-1)/2$, where n is number of the observations. KRCC measure the similarity of data ordering and it is clearly not a true order measurement, since it measures the number of concordant and discordant pairs. KRCC value is between *[-1,1]*, *1* means the agreement between two order is perfect, or both monotonic, otherwise the value is *-1*, while if two measures are independent then the coefficient will be close to *0*.

3.3 Dataset

To see the dynamics of graph pruning effect, we need consistently different graph size that each reflect the real social network characteristics. Currently, the available dataset for real world social network has only approximate social network characteristics and they only have one fixed-size graph. To resolve this problem, we propose to generate several incremental sizes of artificial social network that respect social network characteristics.

The artificial social network generated using *Barabasi – Albert* model [10] that contain scale free distributions and preferential attachment characteristics. There are 8 network, which size range from 50 to 100000 nodes. Each different network is independent from each other, It means that smaller network is not a subset of bigger network or bigger network is not superset of smaller network. Table 1 show social networks size and their properties. We name *Network50* for network with 50 nodes until *Network100000* for network with 100000 nodes.

4 EXPERIMENT AND RESULT

4.1 Single Value Measurement

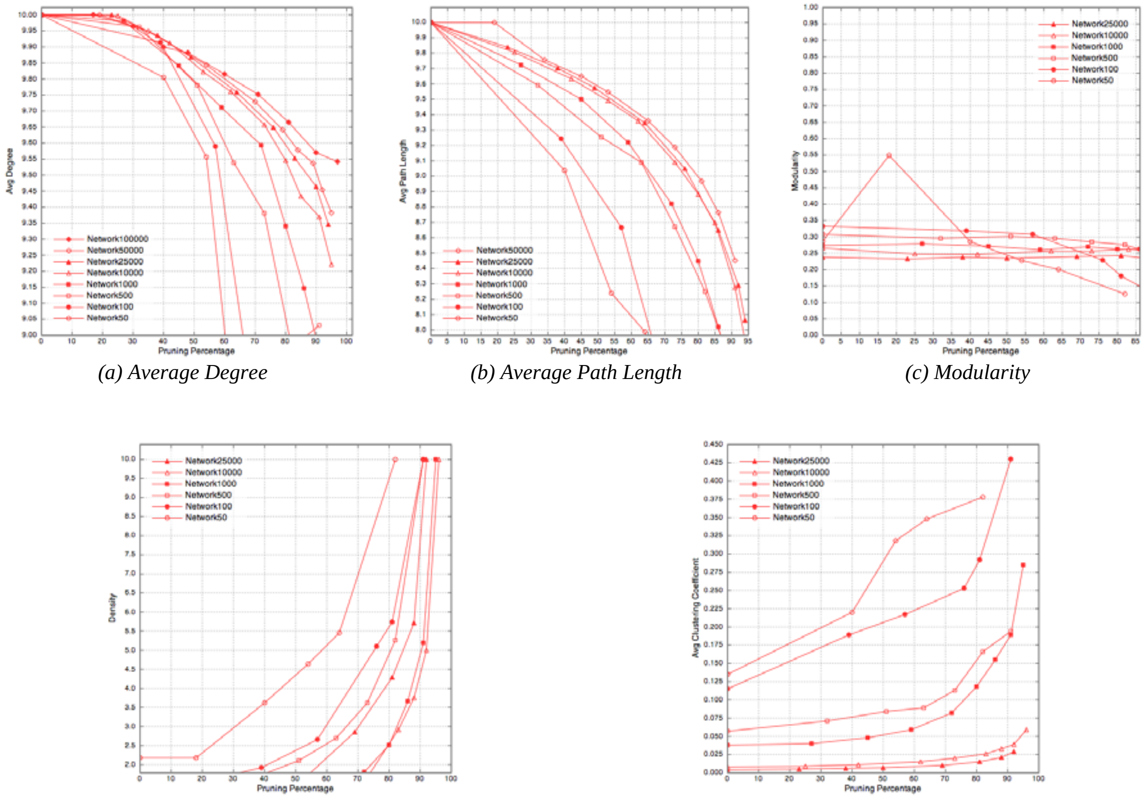
On single value measurement, we plot pruning percentage against properties values. In respect to the evolving graph size, we note some properties behave as monotonic increasing or decreasing function, while some others do not change at all. The value of *Average Degree, Average Path Length, Density,* and *Average Clustering Coefficient* change accordingly in respect to the graph size, while *Modularity* and *Diameter* do not.

We proceed the pruning process based on node degree. We remove node degree below a given threshold. Since node degree distributions is different on each network, then we have different pruning percentage as the result. However, it does

Table 1: different artificial social network size based on Barabasi Albert model generator and their properties

Name	*Nodes*	*Edges*	*AvgDeg**	*APL***	*Diameter*	*Density*	*Modularity*	*ACC****
Network50	50	141	5.6400	2.3069	4	0.1150	0.2870	0.1350
Network100	100	291	5.8200	2.6167	5	0.0590	0.3330	0.1150
Network500	500	1984	7.9360	2.9214	5	0.0160	0.3080	0.0570
Network1000	1000	4975	9.9500	3.0011	5	0.0100	0.2740	0.0380
Network10000	10000	59964	11.9928	3.4982	5	0.0010	0.2660	0.0080
Network25000	25000	174951	13.9961	3.5973	5	0.0010	0.2380	0.0040
Network50000	50000	399936	15.9974	3.6532	5	0.0003	0.2230	0.0027
Network100000	100000	999900	19.9980	3.6429	5	0.0002	0.1970	0.0017

**AvgDeg = Average Degree **APL = Average Path Length ***ACC = Average Clustering Coefficient*



(a) Average Degree

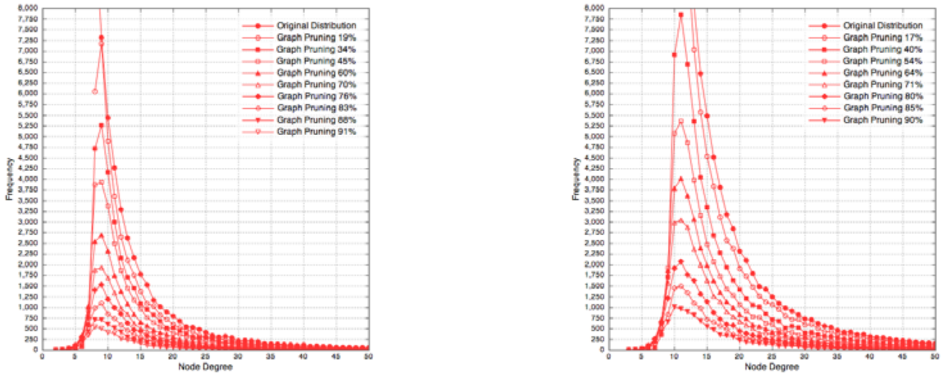
(b) Average Path Length

(c) Modularity

(d) Density

(e) Average Clustering Coefficient

Figure 3. Graph properties evolvement on graph pruning effect in different network size.



(a) Network50000

(b) Network100000

Figure 4. Node Degree Distributions Of Different Pruning Percentage

not hinder our understanding of overall curve trend following pruning percentage.

We plot and normalize properties value of *Average Degree* (Fig. 3a), *Average Path Length* (Fig. 3b), *Density* (Fig. 3d). We plot the original value of *Modularity* (Fig. 3c) and *Average Clustering Coefficient* (Fig. 3e). We do not plot *Diameter*, since mostly the value does not change significantly from the original value for any given sub-graph size. For example, in *Network25000* and *Network50000* for any given pruning percentage the diameter keep on the value of 5 hops.

Some network size in dataset does not present on several properties measurement, especially *Network50000 and Network100000*. Due to their large size, the high time complexity measurement prevents us for computing properties of each pruning percentage within the time frame, but we manage to see curve trend from the consistent value of the smaller network

4.2 Distributions Measurement

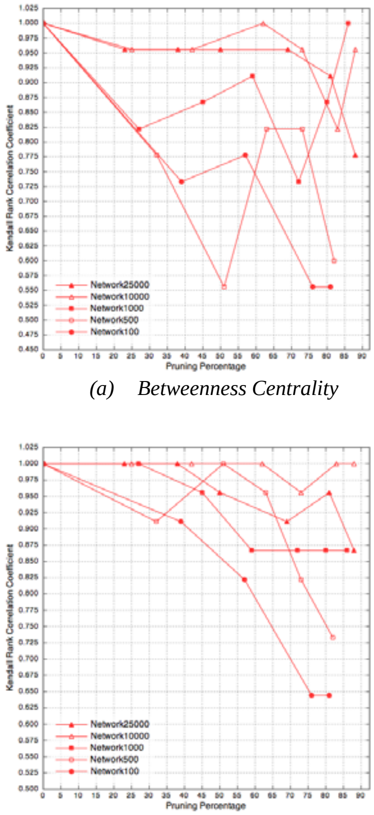
On distribution measurement, we plot *Node Degree Distributions* (NDD) which contain node degree against its appearance frequency in the network. All network NDD behave in similar fashion but in different scale. The scale free characteristics or scale invariants clearly seen on all pruning percentage distributions. The distributions consist of fat-head part and long-tail part, where fat-head contain large number less important nodes and long-tail contain most important node or core part of network. Graph pruning procedure can be illustrated as removal the fat-head part and keep only the long-tail section.

The curve similarity on NDD is measured using FD metrics as explained in 3.2. The accumulation of similarity computation between original graph distribution and each its pruning-step graph NDD indicate the overall similarity distance. Considering that all the curves have similar form, thus the spreading curve area become the main computation factor. As the result, the bigger network has the farther similarity distance. We show only the largest dataset *Network50000* and *Network100000* plot on Fig. 4. as the representatives of NDD measurement, since they have wide spreading area, thus clearer picture comparing to the smaller network.

4.3 Rank Measurement

Investigation on the effect of rank measurement after the pruning process start from the question on how far rank measurement on summary graph has changed from the original rank. Rank distortion can be caused by scenario such as rank-order differences and/or how many step the order change. For example: if we have the original order *1,2,3,4,5,6,7,8*. Given between new rank *3,2,1,4,5,6,7,8* and *3,6,8,1,4,2,5,7*, then the former is more consistent to the original because smaller rank-order differences and fewer order change.

Rank consistency is measured using *Kendall Rank Correlation Coefficient* (KRCC) explained in 3.2. We test *Centrality* metrics based degree, betweenness and closeness. KRCC measures how far correlation between the original rank in each network size and each theirs pruning percentage.



(a) Betweenness Centrality

(b) Closeness Centrality

Figure 5. Kendall Rank Correlation Coefficient Of Centrality Rank Measurement Seen From Different Pruning Percentage

As we prune graph based on node degree, the impact on the degree centrality rank is insignificant, but there are cases where rank is changed because of the particular node lost significant number of connections from the pruning process. However, the different scenario is happened on the case of *Betweenness Centrality* (BC) and *Closeness Centrality* (CC), where betweenness measures shortest path and closeness measures distance. By doing node degree based removal, it will affect the established path and distance measured on the original rank. The overall result of KRCC for dataset tested is the positive correlation, it means that graph pruning can maintain rank order positively in each size tested.

There are no known relations or pattern between conserving both path or distance and node degree removal. Nonetheless, we establish some notes based on the empirical result from the relations between KRCC and pruning percentage in Fig.5. The rules are as follows: (a). smaller network tends to have less consistent rank order as the increasing pruning percentage, we suspect this is because there are few alternatives path and/or distance, so node removal caused BC and/or CC values varies much. (b). some network has consistent rank order, we note that this is because the value gap in the original BB and/or CC is high, otherwise the inconsistent rank order can be caused by smaller value gap, thus after the pruning process, they can easily produce overlap rank.

5 ANALYSIS

On chapter 4, It is shown that the bigger network means they are more robust to properties value change due to pruning process. On smaller network, a similar pruning percentage gives faster property value change. The bigger network also generally more representatives to illustrate graph pruning performance, since the curve are smoother and predicted.

We found around 30-40 percent of graph pruning does not change much single value measurement. The accuracy is above 99% for *Average Degree* and 97% of *Average Path Length*. For *Density* and *Average Clustering Coefficient* property value above *Network500* does not change much until 50-70 pruning percentage. While *Diameter* and *Modularity* value does not significantly change regardless the pruning percentage. As the result, back to graph pruning formulation in 2.3, The properties of *Average Degree, Average Path Length*, *Density* and *Average Clustering Coefficient* can be represented in summary graph or can retain the original graph property value.

Node Degree Distribution on any pruning percentage preserve the original distributions, this clearly fulfill a scale free characteristics. Since both node degree properties, that is *Average Degree* and *Node Degree Distribution* retains very well the original graph properties, then we can use node degree as a base of our graph pruning methods.

There are two reason why we choose node degree: (1). Its simplicity: The cost of constructing NDD is cheap, even for large-scale social network. Node degree rank is very intuitive to simplify combinatorics relationship in social network. Compared to other property such as path length (node betweenness), distance (node closeness), or rank measurement, they are hard to compute and intuitively difficult to formalized in support graph pruning constructions. (2) Its predictability: node degree property scales well as progressing pruning percentage.

We can use graph pruning based on node degree to reduce graph size in very fast fashion, based on *k-core* property shown in 2.2, but the process does not always guarantee can predict accurately other properties like path, distance, and grouping. To predict accurately, we need other complement method to graph pruning, such as graph sampling based random walk after pruning process to accurately preserve network core path, as it has been shown in 2.4.

A stopping criteria is needed for graph pruning process to have an optimal result, a tradeoff between properties accuracy and network size. Other stopping criteria might also have needed such as network connectivity for path measurement or we can state as the number of connected component in the network, we stop pruning process just before network connectivity had gone to preserve path property.

6 CONCLUSION

Our graph pruning approach for reducing large-scale social network complexity is based on node degree context. This idea is based on *k-core* graph property. The graph pruning process requires *scale free distributions* and a stopping criteria value ε to get an acceptable representative result. The pruning methods is very fast to reduce large-scale social network size, while still maintain the original graph properties in context of single value measurement of node degree, density, and path length. However, for rank measurement properties of path and distance, the result might be different from the original, but it gets more accurate as network size getting bigger.

The above result might be different if we use pruning methods based on other graph property context, for example prune based on path rank or distance rank. However, this approach is more complex and expensive compared to node degree context.

For future research, we suggest testing the hybrid approach of graph pruning with other graph sampling methodology to handle large-scale social network. Graph pruning works as a fast size-reducer method and graph sampling works to make sure that sample maintain the original graph properties. In this scenario, graph pruning works as pre-processing step to graph sampling methods.

REFRENCES:

[1] Alamsyah, Andry; Peranginangin, Yahya. “Network Market Analysis using Large-Scale Social Network Conversation of Indonesia’s Fast Food Industry”. *3rd IEEE International Conference on Information and Communication Technology*. 2015

[2] Ward, Michael D; Stovel, Katharina; Sacks, Audrey. “Network Analysis and Political Science”. *Annual Review of Political Science*. Vol 14: 245-264. 2011

[3] Alamsyah, Andry; Peranginangin, Yahya. “Effective Knowledge Management using Big Data and Social Network Analysis. *Learning Organization: Management and Business International Journal*. Vol 1 No 1 ISSN: 2354-6603. 2013

[4] Zhang, Feixiang; Zong, Liyong. “Dissemination of Word of Mouth Based on SNA Centrality Modelling and Power of Actors – An Empirical Analysis Internet Word of Mouth”. *International Journal of Business Adminstration*. Vol 5, No 5. 2014

[5] Alamsyah, Andry; Paryasto, Marisa; Putra, Feriza J; Himmawan, Rizal.” Network Text Analysis to Summarize Online Conversations for Marketing Intelligence Efforts in Telecommunication Industry”. *4th IEEE International Conference on Information and Communication Technology*. 2016

[6] Liu, Bing. *“Sentiment Analysis and Opinion Mining”*. Morgan & Claypool Publishers. 2012

[7] Scott, John. *“Social Network Analysis: A Handbook”*. Sage Publications. 2000

[8] Diestel, Reinhard. “*Graph Theory: Electronic Edition 2005”*. Springer –Verlag Heidelberg, New York 1997, 2000, 2005

[9] Kang, U; Papadimitriou, Spiros; Sun, Jimeng; Tong, Hanghang. “Centralities in Large Networks: Algorithm and Observations”. *Proceedings of SIAM International Conference on Data Mining*. 2011

[10] Albert, Reka; Barabasi, Albert-Laszlo. “Statistical Mechanics of Complex Networks”. *Review of Modern Physics*. 74(1):47-97. 2002

[11] Ahmed, Nesreen K; Neville, Jennifer; Kompella, Ramana. “Network Sampling vie Edge-vased Node Selection with Graph Induction”. *In Purdue University*, CSD TR #11-016, pp: 1-10. 2011

[12] Hu, Pilli; Lau, Wing Cheong. “A Survey and Taxonomy of Graph Sampling”. *CoRR*, abs/1308.5865.2013

[13] Leskovic, Jure; Faloutsos, Christos. “Sampling from Large Graphs”. *Proceeding of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp: 631-636. 2006

[14] Krishnamurty, V; Faloutsos, M; Chrobak, M; Lao, L; Cui, J-H; Percus. A.G. “Reducing Large Internet Topologies for Faster Simulations”. *International Conference on Research in Networking*. pp: 328-341. 2005

[15] Gjoka, Minas; Kurant, Maciej; Butts, Carter T; Makopoulou, Athina.” Walking in Facebook: A Case Study of Unbiased Sampling of OSNs”. *Proceeding of INFOCOM 29th International Conference on Information Communications*. pp 2498-2506. 2010

[16] Hubler, Christian; Kriegel, Hans-Peter; Borgwardt, Karsten; Ghahramani, Zoubin. “Metropolis Algorithms for Representatives Subgraph Sampling”. *8th IEEE International Conference on Data Mining*. 2008

[17] Alvarez-Hamelin, Jose Ignacio; Dall’asta, Luca; Barrat, Alain; Vespignani, Alessandro. “K-Core Decomposition of internet Graphs: Hierarchies, Self-Similarity and Measurement Biases”. *Network and Heterogeneous Media* 3, 371. 2008

[18] Brandes, Ulrich. “A Faster Algorithm for Betweenness Centrality”. *Journal of Mathematical Sociology*. 25(2): 163-177. 2001

[19] Newman, M.E.J. *“Network: An Introduction”*. University Michigan and Santa Fe Institute. Oxford University Press. 2011

[20] Batagelj, Vladimir; Zaversnik, Matjaz . “An O(m) Algorithm for Cores Decomposition of Networks”. *Advances in Data Analytics and Classification*. Vol. 5, No. 2, pp: 129-145. 2003

[21] Elfrat, Alon; Guibas, Leonidas J; Sariel, Har-Peled; Mitchell, Joseph S.B; Murali, T.M. “New Similarity Measures between Polylines with Applications to Morphing and Polygon Sweeping”. *Discrete & Computational Geometry*. Vol. 28, Issue 4, pp: 535-569. 2002

[22] Agresti, Alan. *”Analysis of Ordinal Categorixal Data, 2nd Edition”*. ISBN: 978-0-470-08289-8. New York: Wiley & Sons. 2010