



Reducing Data Social Network Utilizing Greedy Randomized Method

Muhammad Rizqy Alfarisi¹, Andry Fajar Zulkarnain², Andry Alamsyah^{3*}

¹School of Applied Science, Telkom University, Indonesia.

²Faculty of Information Technology, Lambung Mangkurat University, Indonesia.

³School of Economics and Business, Telkom University, Indonesia

¹mrizkyalfarisi@telkomuniversity.ac.id, ²andry.zulkarnain@ulm.ac.id,

³andrya@telkomuniversity.ac.id.

*Corresponding author: ³andrya@telkomuniversity.ac.id.

Abstract

There are many complex data in social networks that can be combined into large data sets which can be accessed. Large amounts of data require more storage and increase the computed summary costs. The need to know how to use data effectively and to extract information from reduced data that has the same information as super data before being reduced. the first thing to do is to convert any unstructured data into structured data utilizing the greedy randomized method, the data can be grouped and combined with other data in its vicinity, and the size of the data can be reduced because the node (user) grouping as a linked pair and is formed the best node around it. This paper presents how to use the minimum description length, as information theory to provide solutions in the model selection problem and apply it in a greedy randomized algorithm that can group unstructured data to reduce data size and provide visualization of the relationship between nodes and how accurate and faster greedy randomized would reduce and combined data into simple link nodes.

Keywords: Social networks, Greedy randomize, Structured data, Unstructured data, Minimum Description Length.

1. Introduction

Social networks are built from a large number of complex data sets, these data sets produce large network sizes, requiring a large server choice to access this data to be faster and more accurate. In this case, it causes problems inefficiency in managing data. Facebook, Twitter, and many popular social networking sites serving millions even billions of users all at once assumed this information from every user as node (user) and edges (their friend list). Mining social media network graphics could provide valuable information about user behavior relationship such as hobbies, what they like, communities, or other activities of users who has similar interests. A common theme on all social networks is a large graph of nodes and edges, millions and even billions. the challenge is to make large graphic visualizations of billions of nodes and edges that can fit into one screen. In addition, developing an algorithm that can scale giant charts is another non-trivial challenge, especially for making graphs that must be compressed into main memory [11]. To overcome that problem, use a greedy randomized method to manage account data on social media such as Facebook and Twitter into structured data and to a grouping that data with specific rules to create simple methods for grouping data and minimizing storage.



2. Research Methodology

2.1. Definition Of MDL

Minimum Description Length is a method for summarizing and giving general solutions to selection model problems, every data that represented in MDL can utilize to reduce another data with a symbol (nodes and edges), observe that MDL theory is how to decode some data into a single "universal" representative model and permits the decoder to construction the code [4].

In the graph, MDL creates a simple connection that represented every problem based on statistical success data even there are have error data. More simple models that could represent more problem is the best, in this method the greater resulting code length is from the closer degree of the data, and the data will be optimal in distribution, in that case, MDL could be easy to utilize in modeling social network graph, MDL representation of the original graph to reduce the data for given a set of minimum data in a graph.

2.2. MDL Representation

$$(R) = |ES| + |C| \quad (1)$$

G represents graph input data, S as summary theory, and C correction which represents data encoding in terms of theory. Defines representation of cost in $R = (S, C)$ as a summary of the storage cost two components, namely, $(R) = |ES| + |C|$, (assume to overlook the cost of stored Av mapping for super $v \in VS$ node because in general, it will be smaller than the storage cost of the ES and C edge sets) [14].

In expressing costs, the first term $|ES|$ according to (A) and second term $|C|$ corresponds to (B). Thus, $R = (S, C)$ shows a representation of minimum cost, then the MDL theory says that S is the best summary of the graph. In another word, R besides being the most compression representation of the graph, also contain S, which is the best graph summary. Refers to a representation of the minimum cost R as a Minimum Description Length represented [9] [10].

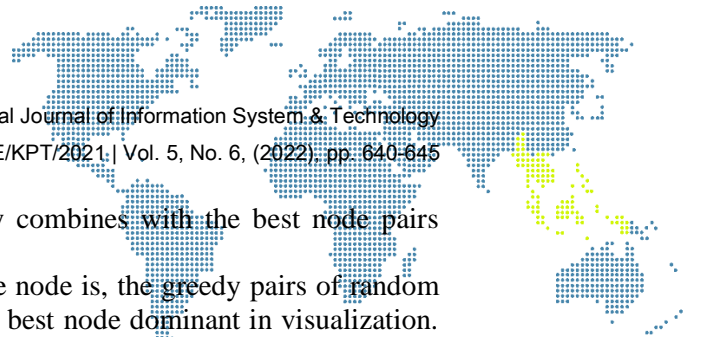
In this case, edges determinate ES and C, the result of representation cost (R), is determined solely based on the super-node which containing in VS. To understand how it is calculated to define VS, assume there are two super-nodes u and v. Define (u, v) as a determinant of pairs (a, b), so $a \in Au$ and $b \in Av$, this determinate represented all possibility G edges that might exist between super-nodes. Next, let $(A u v \subseteq u v)$ become the actual edges that determinate original graph G $(A u v = u v \cap EG)$ [3].

Currently, there are two ways of encoding in $(A u v)$ utilizing correction and summary structures. In the first step, add super edges (u, v) to S and $(u v - A u v)$ edges to represent the negative correction C, second step, add edges to the $(A u v)$ set as a positive correction C. The memory that is needed for these two alternatives are $(1 + |u v - A u v|)$ and $|A u v|$, only select the one which has smaller memory requirements to encode $(A u v)$ edge.

Based on that theory, representing cost of determinate $(A u v)$ edges between super-node u and v is $(c u v = \min \{|u v| - |A u v| + 1, |A u v| \})$. Furthermore, super edge (u, v) would be present in a summary graph of S if $(A u v) > (|u v| + 1)/2$, then selected according to positive and negative corrections. Given VS super-node set the cost of its representation would be calculated each pair of super-nodes and made a simple choice as described above [1].

2.3. Algorithm Representations

Based on Minimum Description Length representation there are two types of algorithms. The first algorithm, Greedy randomized, combines node nodes iteratively with the best nodes that provide maximum cost reduction to become super-nodes. The second



algorithm, Randomized, is a method that randomly combines with the best node pairs globally around it [6] [13].

Optimal data distribution is seen in how close the node is, the greedy pairs of random links would combine from random data to make the best node dominant in visualization. the greedy randomized result will not allow the data reduce into smaller data more than greedy.

2.4. Randomized Algorithm

Randomized is a random scheme which is a very mild random merging procedure. A randomization algorithm is a way to concatenate vertices that are randomly selected around them to create the best node pair, the motive of this scheme is to exchange the computing summary costs to reduce computational complexity and to reduce execution time for fused nodes.

Compared to the Greedy randomized algorithm, it can merge nodes faster, because it randomly chooses every node in the vicinity and combines it with the best node, and allows it to scale very large inputs around it [8].

The Randomized algorithm is not compact as Greedy randomized. Randomized makes it possible to reduce the node to its maximum size as Greedy randomized (note: Randomized cannot reduce the data smaller than Greedy). The random iterative algorithm combines one node to form a set of super-nodes and is divided into two types, (U) stands for un-finished, and (F) stands for finished. The finished type is tracking nodes that do not provide cost reductions with other nodes ((-) negative values for all pairs contained), meanwhile, the unfinished type, includes the remaining nodes which are considered combined by a Random algorithm.

At first, all nodes are unfinished. In every step, randomly select the node from U, represented by u, then found node v, it represents s (u, v), largest pairs that contain u. If combining the nodes provides a positive cost reduction, then combining them into (w) super-nodes. The next steps remove u and v from VS (and U) and add (w) to VS (and U). However, it must combine with another node if it has a negative result [2].

3. Results and Discussion

3.1. Testing Purpose

The purpose of reducing data social networks with the randomized algorithm is [3]:

- To analyze the result of implementation graph reduction data based on a randomized algorithm.
- To analyze the graph data from the stage in the super-node before merging in the vicinity to the node that merges and reduces the data into a small graph.
- To analyze the visualization of those represented from the graph.

In this paper, the system for reducing data is using the MDL method. To analyze and evaluate the performance of the graph model using a randomized algorithm. Sample data for testing is dataset SNAP [15], from the data the reducing process will produce compressed data in a graph model.

Below is SNAP data from Facebook as a sample:

Nodes	Id	Label
236	236	236
252	252	252
276	276	276
26	26	26
280	280	280
272	272	272
84	84	84
133	133	133
62	62	62
315	315	315
200	200	200
67	67	67
21	21	21
248	248	248
25	25	25
13	13	13
30	30	30
257	257	257
303	303	303

Figure 1. Facebook Dataset

For the implementation, it utilizes python as a programming language dan gephy to visualization the pair of merges from the node. Every data processed by a randomized algorithm produce a graph that represents the real graph in comparison with time, compression ratio, and cost.

```

----Connectivity----
236; -> [62, 315, 67, 21, 248, 25, 13, 30, 257, 303, 314, 318, 252, 276, 262, 280, 272, 133, 84, 200]
62 -> [236, 122, 276, 199, 170, 200, 98, 223, 142, 224, 261, 161, 318,]
315 -> [236, 304, 9, 329, 213, 26, 280, 211, 133, 322, 56, 323, 98, 67, 223, 339,
67 -> [236, 186, 285, 271, 213, 252, 26, 280, 272, 62, 322, 188, 119, 323, 313, 325
, 40, 10, 55, 69, 303, 75,]
21 -> [236, 9, 332, 280, 170, 56, 315, 119, 323, 277, 134, 223, 274, 169, 142, 105,
248 -> [236, 332, 26, 53, 119, 323, 200, 67, 277, 21, 169, 265, 158, 38, 311,]
25 -> [236, 186, 285, 176, 9, 252, 26, 280, 272, 199, 84, 56, 231, 119, 323, 200,
, 40, 39, 51, 331, 221, 148,]
13 -> [236, 271, 304, 199, 239, 172, 56, 188, 98, 313, 67, 118, 277, 223, 238, 265,
30 -> [236, 9, 329, 213, 322, 56, 73, 48, 224, 303, 331,]
257 -> [236, 26, 280, 272, 315, 169, 25, 40, 39, 295,]
303 -> [236, 239, 322, 56, 345, 109, 104, 45, 132, 221,]
314 -> [236, 271, 82, 65, 199, 313, 342, 161, 96,]
318 -> [236, 82, 119, 334, 261, 123,]
186; -> [236, 122, 271, 213, 98, 67, 325, 277, 223, 88, 25, 59, 123, 104, 113, 331]

```

Figure 2. Greedy Randomized result from the Facebook dataset

In this stage create the visualization from social network data that reduced and utilizes randomized algorithm, the graph below shows us the differences of the graph before reducing the merge and after utilizing a randomized algorithm.

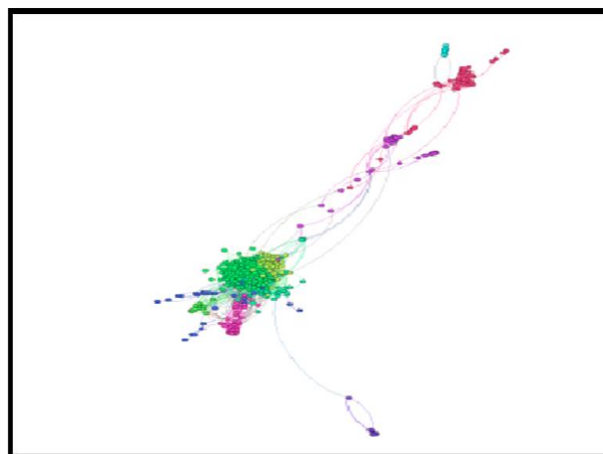


Figure 3. Visualization Original Graph

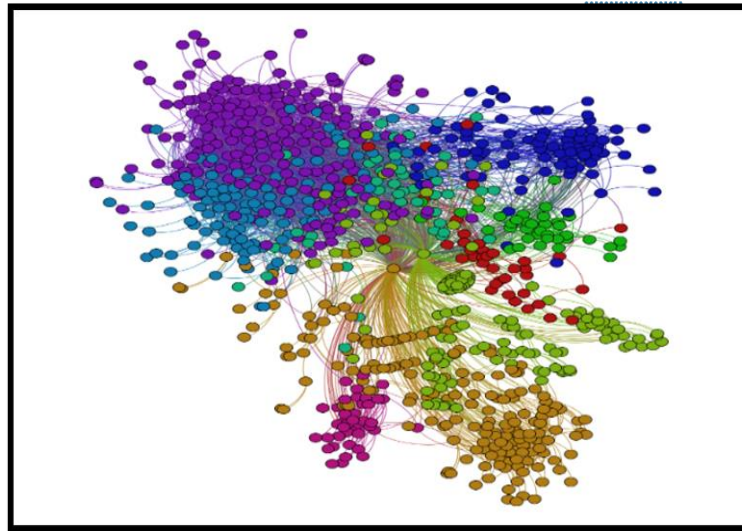


Figure 4. Graph summary after utilizing greedily randomized

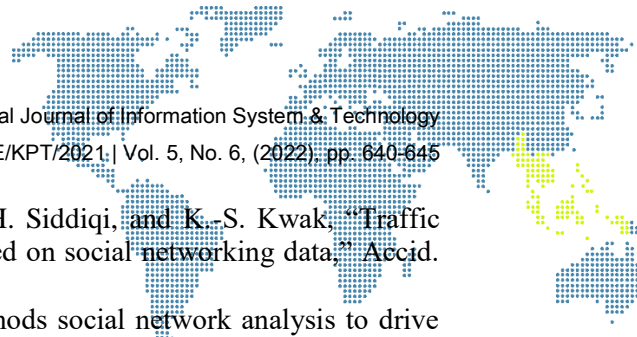
From the experimental data set, the number of edges affects the runtime execution that system needs to create a summary. How long the time that is needed to create the graph summary depends on how many edges on the graph with the big amount of data increase the average degree of the graph. Based on the experiment, the query result in the greedy randomized system has different nodes in neighbor node between the first graph and the graph result on every level of depth from the graph, because there is lost information or unconnected edge to the vicinity in the process of the merge to create a single dominant node.

4. Conclusion

The result from analyzing and testing with the randomized algorithm on social network data. Social networks are a complex problem but they can reduce the visualization using a greedy randomized algorithm and show us the dominant data in the super-node, utilizing minimum description length (MDL) and a greedy randomized algorithm could produce a represented graph that is compact as a greedy algorithm. Reducing data use a randomized algorithm is faster than greedy because sorting data randomly select the node in the vicinity.

References

- [1] Y.Liu and C.-Q. Gu, "Variant of greedy randomized Kaczmarz for ridge regression," *Appl. Numer.Math.*, vol.143, pp. 223-246,2019.
- [2] S. Saurabh, S. Madria, A. Mondal, A. S. Sairam, and S. Mishra, "An analytical model for information gathering and propagation in social networks using random graphs," *Data Knowl. Eng.*, vol. 129, p. 101852, 2020.
- [3] Y. Wu et al., "Identification of subtype-specific biomarkers of clear cell renal cell carcinoma using random forest and greedy algorithm," *Biosystems*, vol. 204, p. 104372, 2021.
- [4] Indonesia Internet Service Provider Association, "Penetrasi & Perilaku Pengguna Internet Indonesia," APJII, Jakarta, 2017.
- [5] Gartner, Inc., "Gartner Hype Cycle: Interpreting Technology Hype," [Online]. Available: <https://www.gartner.com/en/research/methodologies/gartner-hype-cycle>. [Accessed 10 January 2019].

- 
- [6] F. Ali, A. Ali, M. Imran, R. A. Naqvi, M. H. Siddiqi, and K.-S. Kwak, "Traffic accident detection and condition analysis based on social networking data," *Accid. Anal. Prev.*, vol. 151, p. 105973, 2021.
 - [7] T. Palonen and D. E. Froehlich, "Mixed methods social network analysis to drive organizational development," *Mix. methods Soc. Netw. Anal. Theor. Methodol. Learn. Educ.* London Routledge, 2020.
 - [8] J. Stokes and S. Weber, "Common greedy wiring and rewiring heuristics do not guarantee maximum assortative graphs of a given degree," *Inf. Process. Lett.*, vol. 139, pp. 53–59, 2018.
 - [9] M. Boull  , "Hierarchical two-part MDL code for multinomial distributions," *Int. J. Approx. Reason.*, vol. 103, pp. 71–93, 2018.
 - [10] C. Pomare, J. C. Long, K. Churruca, L. A. Ellis, and J. Braithwaite, "Social network research in health care settings: Design and data collection," *Soc. Networks*, 2019.
 - [11] Navlakha, S., Rastogi, R., Shrivastava, N.: Graph summarization with bounded error. In: *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, New York, NY, USA, ACM (2008).
 - [12] Chen, C., Lin, C., Fredrikson, M., Christodorescu, M., Yan, X., Han, J.: Mining graph patterns efficiently via randomized summaries. In: *2009 Int. Conf. on Very Large Data Bases*, Lyon, France, VLDB Endowment (August 2009).
 - [13] Navlakha, S., Schatz, M., Kingsford, C.: Revealing Biological Modules via Graph Summarization. Presented at the RECOMB Systems Biology Satellite Conference. *J. Comp. Bio.* 16 (2009).
 - [14] Alamsyah, Andry. Budi Rahardjo and Kuspriyanto, *Community Detection Methods in Social Network Analysis*, ITB :Bandung, Indonesia.
 - [15] Dataset source: <http://snap.stanford.edu/data/index.html>.