

# Concise Fuzzy Representation of Big Graphs: a Dimensionality Reduction Approach

Faisal N. Abu-Khzam and Rana H. Mouawi

Department of Computer Science and Mathematics  
Lebanese American University  
Beirut, Lebanon

March 9, 2018

## Abstract

The enormous amount of data to be represented using large graphs exceeds in some cases the resources of a conventional computer. Edges in particular can take up a considerable amount of memory as compared to the number of nodes. However, rigorous edge storage might not always be essential to be able to draw the needed conclusions. A similar problem takes records with many variables and attempts to extract the most discernible features. It is said that the “dimension” of this data is reduced. Following an approach with the same objective in mind, we can map a graph representation to a  $k$ -dimensional space and answer queries of neighboring nodes by measuring Euclidean distances. The accuracy of our answers would decrease but would be compensated for by fuzzy logic which gives an idea about the likelihood of error. This method allows for reasonable representation in memory while maintaining a fair amount of useful information. Promising preliminary results are obtained and reported by testing the proposed approach on a number of Facebook graphs.

## 1 Introduction

Graphs can represent relations that exist in several domains. Social networks are a typical example of such usage with users as nodes that connect to one another by edges. The constant growth of such networks and the corresponding immense size of graph representation impose some difficulties

in maintaining and manipulating such data. In particular, the edges that make up a graph can exceed its corresponding nodes by a great margin so that representation using an adjacency list is nearly impossible when dealing with big graphs. This is especially true when using a system with a conventional main memory.

Notable examples of big graphs include graphs that are used to represent webpages and their interconnections (i.e., hyperlinks) which can have 20 billion nodes and 160 billion edges. Social networks, on the other hand, can contain a billion users with relationships among users reaching more than 140 billion. If the graph were to be maintained solely on disk, i.e. using virtual memory, then the I/O operations required to answer queries concerning it would take up too much computing time as opposed to queries that consult a main memory.

Our aim in this work is to transform a typical graph representation into a more compact form. Currently, several frameworks exist to deal with sizeable graphs such as Pregel [7], Apache Giraph [10], and GraphLab [6]. These projects are based on leveraging the computing resources of many machines so that the operations on the graphs are distributed. A recent approach for graph compression, on the other hand, relies on identifying repeated patterns in graphs and representing them through “grammar rules” [8]. This approach aims to enhance the performance of certain types of queries.

Our proposed approach is based on the following: When extracting useful information from graphs, it is possible that not all edges are of the utmost importance to reach a conclusion. That is, some inaccuracy might be tolerated during analysis. This is especially the case if we are able to get an estimate of whether or not two nodes are neighbors. The result is that we find a compromise between maintaining all data in a graph and being able to store that graph in a typical main memory. Our approach moves us from needing a quadratic amount of storage in the number of nodes to a linear amount. This is further detailed in the next section.

## 2 A Graph Mapping Approach

In order to manage the large amount of data associated with graphs, we seek to transform their representation to a more compact form, even if some information is lost. Our proposed method is to represent the nodes in a  $k$ -dimensional space, where  $k$  is a fixed constant, with distances between them indicating their adjacency status. This is done by associating each node with two parameters,  $r$  and  $R$ , that indicate the “radius” within which its

neighbors are located and the “radius” outside of which its non-neighbors are located, respectively. There is uncertainty regarding the nodes that lie within these two values.

For the first part, we propose to use the linear-time FastMap algorithm from [5], which takes a distance matrix and returns a set of points in a  $k$  dimensional space where  $k$  is user defined. FastMap works in linear time in the number of nodes thus it outperforms other algorithms with similar goals such as multidimensional scaling. Once we have this mapping, it is possible to calculate the two parameters of each node  $v$ . This is done by calculating the distance from  $v$  to all other nodes, and arranging those distances, for example, in an increasing manner. Iterating from the beginning of this list, we can determine the least “ $r$ ” such that all nodes within it are neighbors of  $v$ . Similarly, from the end of this list we may conclude the value of “ $R$ ”. With this approach, of course, the resulting representation will not give a definitive “yes” or “no” about the adjacency status of two nodes unless it is truly the case.

When the Euclidean distance  $d$  between two nodes  $v$  and  $v'$  lies within the values of  $r$  and  $R$ , we invoke a fuzzy logic system that gives us the likelihood of those two nodes being neighbors, with larger outputs indicating a higher chance of them being adjacent. This system assumes the closer  $d$  is to the value of  $r$  as compared to  $R$ , the more likely it is that  $v$  and  $v'$  are adjacent. The opposite is assumed when the measured distance is closer to the value of  $R$ .

The system takes as input a crisp value between 0 and 1 that is obtained by dividing the difference of  $R$  and  $d$  by the difference of  $R$  and  $r$ . The inputs are subjected to two membership functions that correspond to describing how close the value is to each of  $r$  and  $R$ . The inference system relies on two rules: if the input value is “close to smaller  $r$ ”, then the two nodes are neighbors and if the input value is “close to larger  $R$ ”, then the two nodes are not neighbors. These rules are specified in an “fcl” file in a fuzzy control language. [3, 4] There are also two membership functions for the output set that correspond to whether two nodes are adjacent or non-adjacent. The activator of the inference rules is the minimum operator which truncates the output membership functions for each rule. The functions of the input and output sets are represented below.

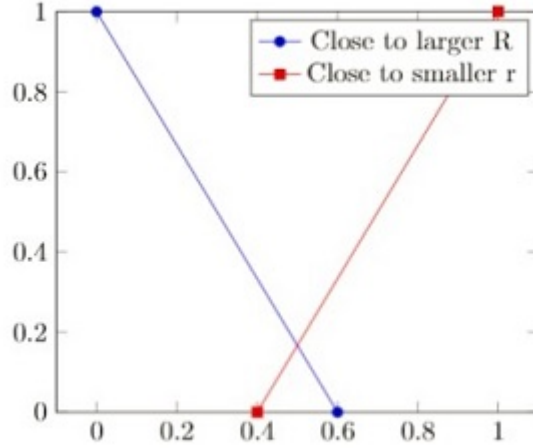


Figure 1: Input Membership Functions.

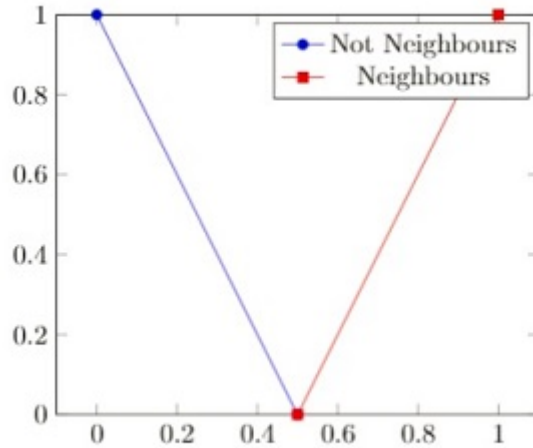


Figure 2: Output Membership Functions.

Accumulation of the inference rules' results takes place using the maximum operator. Finally, defuzzifying the output value is done using the center of gravity method.

Our algorithm starts by running FastMap, which outputs the coordinates of the vertices in a  $k$ -dimensional space. Since we are mainly interested in unweighted graphs, especially those representing social networks, the distance matrix is built by setting a distance of one between neighboring nodes and a distance of  $n$  otherwise, where  $n$  is the number of nodes/vertices in

the graph. The values of  $r$  and  $R$  for all nodes are calculated as integer values as described earlier. To determine whether two vertices  $v$  and  $v'$  are adjacent, the Euclidean distance  $d$  between them is measured and then it is compared according to the following algorithm.

---

**Algorithm 1** Adjacency Query

---

```

1: if  $d \leq r(v)$  or  $d \leq r(v')$  then
2:   return 1
3: else if  $d \geq R(v)$  or  $d \geq R(v')$  then
4:   return 0
5: end if
6:  $input1 = (R(v) - d)/(R(v) - r(v))$ 
7:  $output1 = \text{invokeFuzzyLogicSystem}(input1)$ 
8:  $input2 = (R(v') - d)/(R(v') - r(v'))$ 
9:  $output2 = \text{invokeFuzzyLogicSystem}(input2)$ 
10: return  $\text{minimum}(output1, output2)$ 

```

---

An example of our approach can be observed in the following two figures.

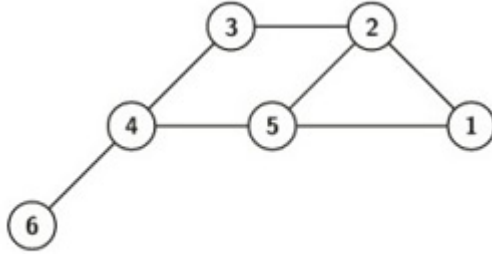


Figure 3: A sample input graph.

Each vertex in the graph will be mapped to a point in a 2-dimensional space using FastMap, yielding the following alternate representation. The values for  $r$  and  $R$  are also calculated for each node.

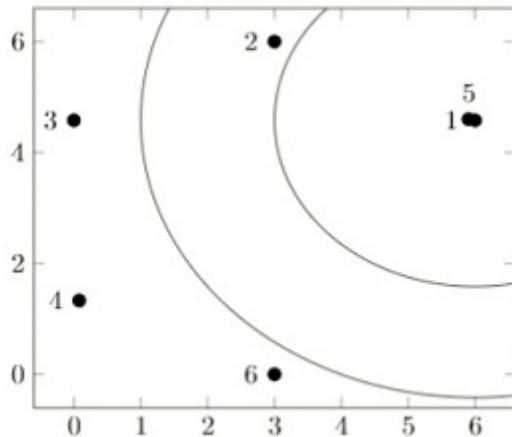


Figure 4: Mapping of graph nodes to 2-dimensional space.

The two circles that correspond to node 1 are shown in Figure 2. Node 5 lies within the smaller circle, which indicates that nodes 1 and 5 are definitely neighbors. The non-neighbors of node 1, nodes 3, 4 and 6, lie strictly outside the larger circle. There is uncertainty regarding node 2. However, as it is closer to the smaller circle, the fuzzy inference system will report a higher chance of the two nodes being neighbors.

Finally, we note that our approach works for directed graphs as well. The query in this case takes an ordered pair  $(v, v')$  as input and the condition for a yes-answer would simply depend on  $r(v)$  and  $R(v)$  only. In fact, the condition  $d \leq r(v)$  would be enough to conclude that there is an arc (or directed edge) from  $v$  to  $v'$ . Moreover, in the case of uncertainty, we would only compute the values of *input1*, then *output1* would be returned.

### 3 Experiments

Experiments were conducted on a number of Facebook graphs, obtained from [9]. The following graph shows the percentage of definite answers maintained after compression as a function of  $k$ . The value of  $k$  must be as small as possible so as to guarantee an effective compression.

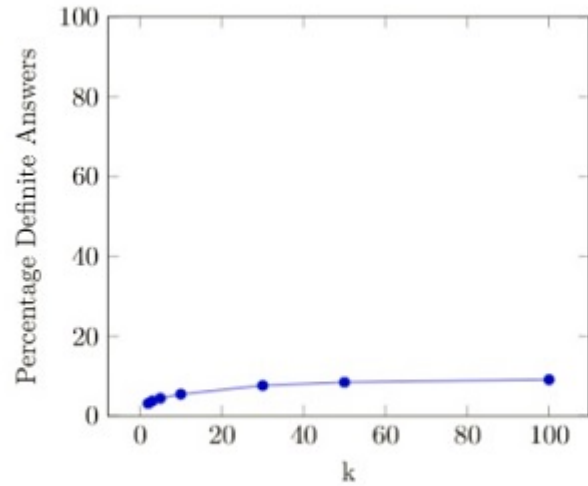


Figure 5: Percentage of Definite Answers Maintained after Compression.

While the above results suggest a tremendous loss of information, fuzzy answers can make up for more certain ones. A fuzzy answer is considered “sound” if it returns a value above 0.5 for a pair of nodes that happen to be neighbors and a value below 0.5 for a pair that are non-neighbors. The figure below shows, among fuzzy queries, the percentage of sound “yes” and sound “no” answers as a function of  $k$ .

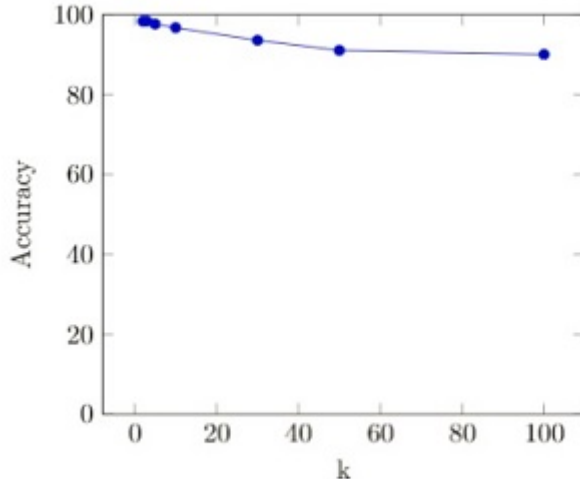


Figure 6: Accuracy of Fuzzy Answers as a function of  $k$ .

The above results reveal a somewhat counterintuitive finding: lower dimensions are required to maintain a higher accuracy of fuzzy answers. A preliminary speculation would predict that higher values of  $k$  offer a higher level of accuracy, as we are reducing the amount of compression. However, the observed behavior could be related to the distribution of data points in a higher dimensional space. Euclidean distances in such spaces might take on differing interpretations than typical 2-dimensional and 3-dimensional distances.

## 4 Conclusion and Future Work

We have proposed a method of compressing graphs that makes up for lost information through fuzzy logic. The compression maps a set of nodes with interconnecting edges to a set of points in a  $k$ -dimensional space. The distance between a pair of points is to indicate whether or not they were connected by an edge in the original graph. Our testing showed a significant percentage of accuracy in the resulting fuzzy representation.

Our reported results describe work in progress and are thus preliminary. An unusual finding that is worth exploring was the decrease in accuracy of fuzzy queries with higher values of  $k$ . This could be attributed to the fact that the distance metric used does not behave according to expectations in higher dimensional settings (see [2], for example). For future work, we



may vary the distance measure used and compare the performance of our approach accordingly.

When dealing with very large graphs or networks, data might be initially stored on several machines. In such settings, the distributed FastMap approach described in [1] can be used. In addition, further testing can be performed while varying the parameters of the fuzzy inference system. As the latter relies on approximate measures, it can be the case that a different membership function or operator could better suit our approach and needs to be determined by thorough experimentation.

## References

- [1] F. N. Abu-Khzam, N. F. Samatova, G. Ostrouchov, M. A. Langston, and A. Geist. Distributed dimension reduction algorithms for widely dispersed data. In S. G. Akl and T. F. Gonzalez, editors, *International Conference on Parallel and Distributed Computing Systems, PDCS 2002, November 4-6, 2002, Cambridge, USA*, pages 167–174. IASTED/ACTA Press, 2002.
- [2] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional spaces. In *Proceedings of the 8th International Conference on Database Theory, ICDT '01*, pages 420–434, London, UK, UK, 2001. Springer-Verlag.
- [3] P. Cingolani and J. Alcalá-Fdez. jfuzzylogic: a java library to design fuzzy logic controllers according to the standard for fuzzy control programming. In *International Journal of Computational Intelligence Systems*, pages 61–75. IEEE, 2013.
- [4] P. Cingolani and J. Alcalá-Fdez. jfuzzylogic: a robust and flexible fuzzy-logic inference system language implementation. In *FUZZ-IEEE*, pages 1–8. IEEE, 2012.
- [5] C. Faloutsos and K.-I. Lin. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In M. J. Carey and D. A. Schneider, editors, *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, pages 163–174, San Jose, California, 22–25 1995.
- [6] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, and J. M. Hellerstein. Distributed graphlab: A framework for machine learning

and data mining in the cloud. *Proc. VLDB Endow.*, 5(8):716–727, Apr. 2012.

- [7] G. Malewicz, M. H. Austern, A. J. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski. Pregel: A system for large-scale graph processing. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 135–146, New York, NY, USA, 2010. ACM.
- [8] S. Maneth and F. Peternek. Grammar-based graph compression. *CoRR*, abs/1704.05254, 2017.
- [9] R. A. Rossi and N. K. Ahmed. The network data repository with interactive graph analytics and visualization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- [10] S. Sakr, F. M. Orakzai, I. Abdelaziz, and Z. Khayyat. *Large-Scale Graph Processing Using Apache Giraph*. Springer Publishing Company, Incorporated, 1st edition, 2017.