NETWORK(GRAPH) DATA RESEARCH IN THE COORDINATE SYSTEM

LIU HUI
College of Information Science and Technology, Beijing Normal University
Beijing 100875, China
School of Computer Science, Anhui University of Technology
Maanshan 243002, China

LIN ZHI* AND WAQAS AHMAD
College of Information Science and Technology, Beijing Normal University
Beijing 100875, China

(Communicated by Zhipeng Cai)

ABSTRACT. Many approaches have been proposed to perform the analysis of network(graph) data correlated with Internet, social networks, communication networks, knowledge graph etc, but few of them have been applied in the coordinate system which is thought to be an efficient platform to processing graph data. In this survey, we provide a short yet structured analysis of network(graph) data research in the coordinate system. We first introduce the coordinate embedding and transforming of double-loop network with its symmetrical and regular structure. We then present two categories of approaches for Internet embedding in the coordinate system and the technology of graph embedding of social network. Finally, we draw our conclusions and discuss potential applications and future research direction.

1. Introduction. With the advent of big data, people not only concern about the applications of data, but also emphasize on data analysis. Application requirements are no longer limited to transactional operations, but on how to effectively obtain valuable information from the data. Thus the relevance of data is drawing more attention.

Graph is a natural tool for studying the relevance of data, used to represent information in various areas [9], including gene interaction networks in bio-informatics field, social networks, semantic networks in linguistics, knowledge graph in search etc. Recently with the rapid development of Internet of Things and 5G wireless networks, the analysis of graph data is in great demand.

Nowadays, information grows explosively, which leads to the increasing scale of the graph. Large-scaled graphs are used in a variety of applications. How to effectively analyze and query large-scale graph scaling well with graph size is a rising hot research subject in academia and in industry.

2010 Mathematics Subject Classification. Primary: 58F15, 58F17; Secondary: 53C35.
Key words and phrases. Coordinate embedding, coordinate system, graph, social networks.
* Corresponding author: linzhi.
1.1. **Challenges.** With the increasing expansion of graph data, there are billions of nodes and edges in the graph. The traditional method, due to high time and space costs, is not applicable to large-scaled graph.

The more common approach is to use the simple structure of the graph [1][2][12], such as the link (path) or tree structure to pre-calculate the time and space costs. While these optimized queries can improve performance locally, there is still a need for global improvement on the large-scaled graphs. In addition, there are some methods dealing with large-scaled graph of a particular structure, such as the compressed BFS-Tree method [23]. This method utilizes the symmetry characteristics of the graph data to optimize the design of graph query and cannot be used to the graphs with little symmetry. Therefore, we need new and efficient ways to deal with the query in big graphs.

Referring to graph, we naturally relate it to coordinate system since the coordinate system sets up a bridge between algebra and geometry therefore, Geometric problems can be described by algebraic methods. For large-scaled graph data, the most efficient way to process each node is using the coordinates. Once the large-scaled graph data is successfully embedded in the spatial coordinate system, the response speed of the query can be greatly improved and the query is not affected by the size of the graph, according to the coordinates of each node.

How to map nodes in high dimensional graphs to positions in low dimension coordinate spaces? To answer this question, We can make an overview of network(graph) data research in the coordinate system in recent years.

1.2. **Organization of the survey.** The survey is organized as follows: In Section 2, we start from double-loop network (DLN) to illustrate coordinate embedding and transforming in the coordinate system. In Section 3, we present two categories of coordinate system of Internet, mapping Internet hosts to a specific position in a Euclidean space, and in Section 4, we propose graph coordinate system, which maps nodes in high dimensional graphs to positions in low-dimension coordinate spaces. At last, we summarize our results and make some concluding remarks.

1.3. **Our contribution.** This survey provides a two-pronged contribution:

1. We propose network(graph) data research in the coordinate system and believe it is a promising research area.
2. To foster further research in this topic, we give the future research directions in the field of coordinate embedding of big graph.

To the best of our knowledge, this is the first paper to survey network(graph) data research in the coordinate system in recent years.

2. **Double loop network.** In the early stage of computer network, the most common topology is ring. The ring has a large diameter and a high vulnerability to node and/or link failures. The double-loop network (DLN) can be viewed as an improvement of the ring topology that overcomes the above mentioned drawbacks.

A double-loop network (DLN) is a widely used interconnection topology in local area networks, for its simplicity, symmetry and scalability. DLN has been categorized into two groups: directed DLN (also named as unidirectional DLN) and undirected DLN (bidirectional DLN). A directed DLN $G(N;r,s)$ ($N \geq 4$) is a directed graph of $N$ nodes which are noted as 0,1,2,...,$n$. For any node $u$, it is connected to nodes $u+r$ (mod $N$) and $u+s$ (mod $N$) by an “r link” and an “s link” respectively. Unlike directed DLN, node in undirected DLN has two more interconnected nodes.
u-r(mod N) and u-s(mod N) except for u+r(mod N) and u+s(mod N). A DLN is strongly connected if and only if N, r and s are relatively prime. Figure 1 shows directed DLN G(8;2,3).

![Directed DLN G(8; 2,3)](image)

The DLN has been studied extensively over the past 30 years since it was introduced in [22]. In particular, it has been studied with respect to network properties such as diameter [10], fault tolerance [6][2], and message routing [8][27]. Currently, though DLN is no longer a hot spot of computer network, the research methodology concerning DLN is still of significance to reference.

2.1. **Coordinates embedding.** A DLN G(N;r,s) can be embedded into Cartesian coordinates because of its symmetrical and regular structure. The source node 0 is considered to be the origin of Cartesian coordinates, where the X axis is represented by r links and the Y axis is represented by s links. The location (x,y) on the coordinates is occupied by node xr+ ys(mod N). Figure 2 shows its embedding graph. The region containing all the optimal routes from a given source node to the other nodes is the MDD(Minimum Distance Diagram) of a BDLN, Which is just an “L-shaped” region shown as the shaded region of Figure 2. An L-shaped tile of BDLN can be characterized by six parameters a, b, m, n, p, q (4 of them independent, \(a = p + n, b = m + q\)). In Figure 2, \(p = 1, m = 2, n = 2, q = 2, a = p + n = 3, b = m + q = 4\).

![embedding graph of G(8;2,3)](image)
Four parameters including \(a, b, m\) and \(n\) can be calculated by following equations [14]:

\[
\begin{align*}
b &= \min k \mid ks \equiv jr(modn), k \leq j \leq 0, k = 1, 2, \ldots, N - 1 \\
m &= j \mid bs \equiv jr(modN), j \leq 0 \\
a &= \min j \mid ks \equiv jr(modN), j > k \leq 0, j = 1, 2, \ldots, N - 1 \\
n &= k \mid ks \equiv ar(modN)
\end{align*}
\]

Thus, Diameter of a directed DLN \(G(N;r,s)\) is \(a + b - \min(m, n) - 2\). Figure 2 shows \(G(8;2,3)\) with a diameter of 4.

2.2. Coordinates transforming. Since a DLN is vertex transitive, any node can be easily relocated in the coordinates. Take node 0 for an example, it can be located in any of the four quadrants on the XY plane, using these equations:

\[
\begin{align*}
p + qh &= 0(modN), -m + bh = 0(modN), a - nh = 0(modN)
\end{align*}
\]

In the first quadrant on the XY plane, the nearest coordinate of 0 is \((p,q)\) and the fourth quadrant \((a,-n)\). Define \(\rho\) and \(\sigma\) as vectors from the origin to the copies of node 0 at \((p,q)\) and \((a,-n)\). Thus, vectors from any node to its four quadrant equivalent nodes can be made respectively with \(\rho = px + qy\) and \(\sigma = ax - ny\).

Embedding a DLN into coordinates system brings intuition and convenience to the research of DLN, especially to the network properties such as diameter and optimal message routing. By means of coordinates transforming, any node can be relocated through vector \(\rho\) and \(\sigma\) or combination of them. With these relocations, BDLN and UDLN turn out to be associated with each other. Figure 4 shows MDD of \(G(39;1,17)\) in the first quadrant, which is a “L-shaped”, and the shaded region shows MDD of \(G(39;\pm 1,\pm 17)\).

Coordinates embedding and transforming even make parallel routing (fault-tolerant routing) and wide diameter (fault-tolerant diameter) calculating of DLN become easily, once wide diameter calculating thought to be NP hard for networks. As illustrated in Figure 5, nodes on axes can be relocated in the areas D1 to D4, visualization of this is of great benefit to the research of parallel routing (fault-tolerant routing) of DLN. For more details about this, we refer readers to [10].

For more information of double-loop networks and their applications, we refer readers to the survey paper written by Hwang [11].
3. **Network coordinate system.** In recent years, large-scaled distributed Internet system has been very widely used. In these systems, there may be thousands of Internet nodes to collaborate, through the interaction of these network nodes to achieve collaborative computing and information sharing. If the distance between the network nodes can be predicted quickly, the overall performance of the network application will be improved greatly.

Network coordinate systems [4] [5] [7] [13] [20] [16] were designed as efficient and salable mechanisms to estimated distances or latencies between Internet hosts. As illustrated in Figure 6, the key idea of GNP is to model the Internet as a geometric space (e.g. a 3-dimensional Euclidean space) and characterize the position of any host in the Internet by a point with a coordinate based on round-trip measurements to other hosts. Once a pair of nodes has converged to their positions in the coordinate space, their network distance in the Internet can be predicted by computing the geometric distance between them. Such distance estimation mechanisms can prove critical to large-scale distributed systems that use approximate distance values for performance optimization.
Based on the way coordinates are computed for new nodes, Network coordinate systems can be generally categorized into "landmark-based" and "decentralized" systems.

3.1. **Landmark-based systems.** Global Network Positioning (GNP) [20], one of landmark-based systems, is the first set of Network coordinate system proposed in 2002, where nodes are divided into two parts, the landmarks and the rest ordinary nodes.

GNP chooses k landmark nodes from total N nodes, to map them into d-dimensional geometric space, $k > d + 1$. GNP then calculates coordinates of k landmark nodes using pair-wise measurements, where errors between measured network distances and geometric coordinate distances are minimized using a non-linear optimization algorithm such as Simplex Downhill [15].

Given the k landmarks coordinates, GNP can next compute the coordinate of any node A among the rest ordinary nodes based on the measured latencies between A and each of the landmarks. Node A computes its own coordinate so that errors of network distances and geometric coordinate distances between A and each landmark are as minimized as possible. This is again achieved by means of the Simplex Downhill method.

Landmark-based systems have fast convergence properties, since all nodes rely on the same fixed nodes for their coordinate calculations. However, because of the need to deploy central servers which known as landmarks, the central servers need to bear the heavy measurement load when the number of participating nodes is large, so the size of the system service is limited and the accuracy of the system may suffer if the choice of landmark nodes is sub-optimal, i.e. they do not sufficiently cover the network.

3.2. **Decentralized-based system.** In contrast, decentralized Network coordinate systems such as PIC [4] and Vivaldi [5] allow incoming nodes to orient themselves in the coordinate space using any nodes already positioned in the space.

Vivaldi regards the entire Internet as a spring system. It is assumed that there exists dummy springs that connect all the nodes in the Internet. Each node acts as a spring against each other to form the entire network. The magnitude of the force of spring is proportional to the difference between measured network distances and predicted coordinate distances. When the actual measured distance and the predicted distance of the nodes are not equal, the elastic potential energy of the nodes can be adjusted to minimize the elastic potential energy of the entire system.
thereby reducing the error of the actual distance and the predicted distance of the nodes.

In the Vivaldi system, each node has its own network coordinates and local error. All nodes periodically update their own network coordinates and local errors according to their measurements on other nodes in the network and the network coordinates of the measured nodes.

While these systems avoid dependence on well-known landmarks, new nodes can force already calibrated nodes to adjust their coordinates, potentially increasing convergence time and propagating errors [13][16]. Though Network coordinate systems has been widely deployed in many well-known applications such as SBON, Bamboo DHT and Bit Torrent for its simplicity and high performance, it has several limitations. Triangle inequality violations (TIV) may be a major barrier for the accuracy of such systems. As illustrated in Figure 7, in the real network, \( AB > AC + BC \), but it does not happen in a triangle.

![Figure 7. triangle inequality violations (TIV)](image)

For further details on Network coordinate systems, we refer readers to a survey [7].

4. **Graph coordinate system.** Since the network nodes on the Internet have triangle inequality violations (TIV), the accuracy of mapping the nodes to the spatial coordinate system is limited. Unlike latencies between Internet hosts, however, shortest path values on a graph, will never violate the triangle inequality. Graph Coordinate System was first proposed in [24], mapping nodes in high dimensional graphs to positions in a fixed dimension Euclidean coordinate space. It is one of the most efficient methods to query the shortest path between different nodes of graph, once a graph is embedded, shortest path queries can be resolved in constant time using the embedded coordinates, i.e. \( O(1) \), independent of the size of the graph. Figure 8 shows mapping graph nodes into Euclidean coordinate space. The shortest path between node A and E is 3 hops in the left graph and the estimated Euclidean distance between them is 2.9 hops in the right graph.

After embedded into coordinate system, each node on the graph has a coordinate, and the shortest distance can be quickly calculated between nodes. This is attractive for applications like network centrality computation or distance based calculating such as information dissemination, community detection, neighborhood function, each of which relies on resolving a large number of shortest path distance queries [24][25][26][18].

Similar to centralized landmark-based network coordinate systems (GNP), nodes in large-scaled graph are embedded into coordinate systems in two phase. The first mapping nodes which are more active called landmarks, and then the rest nodes
are embedded using the distance from themselves to the landmarks. If the mapping accuracy of landmarks is not high, the errors will be further amplified in the mapping of the rest nodes, which may seriously affect the mapping of the whole graph.

Compared to the rest nodes, landmarks are important nodes, which can quickly affect the rest nodes in the whole graph. Intuitively, high degree nodes reside at the core of graph, were chosen as landmarks in [24]. Except for degree, the metrics of the importance of the nodes also be measured by centrality [19], semi-local centrality [3], betweenness centrality [17], etc. How to improve the accuracy of landmark selection based on the importance of nodes represented by the metrics? Landmark selection strategy in [18] is based on local betweenness indicator in sub-graphs instead in the whole graph, which is turned out to be more effective in large-scaled graph with billion nodes.

The number of landmarks should be more than a certain value, and landmarks should be evenly distributed as much as possible in the graph. If the number of landmarks is too small or the distribution is too dense in the local area, the skeleton of the mapping system is not strong, which will lead to the difficulty of mapping the rest nodes, on the contrary, if the number of landmarks is too large, the complexity of the mapping will be increased dramatically. If the uniformity of landmarks is not enough, it is also difficult to guarantee the mapping accuracy of the rest nodes.

The literature [25] proposed Rigel, mapping large-scaled graph nodes to the Hyperbolic coordinate space, which is used in both [25] and [18]. Another way to improve efficiency and effectiveness of graph embedding is to distribute whole graph over a set of machines by partitioning the graph into many sub-graphs, thus allowing parallel process to achieve high speedup [25] [18]. Figure 9 shows the main method of a coordinate-based embedding for large-scaled graph [18], it consists of two parts: offline distance oracle construction (steps S1-S5) and online shortest distance query answering (step S6).

5. **Conclusion and future work.** In this paper, we have a view of network(graph) data research in the coordinate system. After introducing the coordinate embedding and transforming of double-loop network, We present two categories of approaches for Internet embedding in the coordinate system and the technology of graph embedding of social network data. We believe there are five promising research directions on network(graph) data research in the coordinate system:

(1) Network(graph) partitioning. Since the large-scale graph is difficult to process and analyse, dividing the large-scale network(graph) into a certain number of sub-networks(graphs) by clustering or segmentation may be a good way. Thus, we can process and analyse each sub-network(graph) using appropriate strategy.
(2) Error optimization algorithm. The optimization algorithm used in Network coordinate systems and Graph coordinate systems named Simplex Downhill is to optimize errors between virtual and coordinate distance. How to choose other appropriate optimization algorithm and coordinate system to adapt for different networks is of importance to future research.

(3) Landmark selection. The quality and the quantity of landmark nodes are essential for improving the accuracy of the mapping which will improve the accuracy of the query. How to select a certain number of landmark nodes in the large-scaled graph with uniform distribution is a future research direction.

(4) Combination with other methods. The growing research on deep learning has led to some deep neural networks based methods applied to graphs. The research direction is to combine these methods with coordinate embedding and transforming of graph for dimensionality reduction.

(5) So far the large-scaled graphs we handle are mostly related to social networks and complex networks. How to extend the processing method of the graph in the coordinate system to other networks(graphs) in order to enhance the extensibility of the processing method needs further study.

REFERENCES

[1] C. C. Aggarwal and H. Wang, Managing and Mining Graph Data, Springer-Verlag, New York, 2010.
[2] Y. Chen, Y. Li and T. Chen, Optimal fault-tolerant routing algorithm and fault-tolerant diameter in directed double-loop networks, Theoretical Computer Science., 468 (2013), 50–58.
[3] D. B. Chen, L. Lu and M. S. Shang, Identifying influential nodes in complex networks, Physica A., (2012), 1777–1787.
[4] M. Costa, M. Castro, A. Rowstron and P. Key, Pic:Practical internet coordinates for distance estimation, In Proc. of ICDCS., (2004), 178–187.
[5] F. Dabek, R. Cox, F. Kaashoek and R. Morris, Vivaldi: A decentralized network coordinate system, In Proc. of SIGCOMM., (2004), 15–26.
[6] H. P. Dharmasena and X. Yan, An optimal fault-tolerant routing algorithm for weighted bidirectional double-loop networks, *IEEE Trans on Parallel and Distributed Systems*, 16 (2005), 841–852.

[7] B. Donnet, B. Gueye and M. A. Kaafar, A survey on network coordinates systems, design, and security, *IEEE Communication Surveys and Tutorials*, 12 (2010), 488–503.

[8] A. A. Farrag, Developing fault-tolerant distributed loops, *Information Processing Letters*, 111 (2010), 97–101.

[9] P. Goyal and E. Ferrara, *Graph Embedding Techniques, Applications, and Performance: A Survey*, IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017.

[10] H. Liu, Z. Zhang and M. Y. Fang, Research on optimal parallel routing and wide diameter of unidirectional double-loop networks, *Journal on Communications*, 8 (2014), 63–70.

[11] F. K. Hwang, A complementary survey on double-loop networks, *Theoretical Computer Science*, 263 (2001), 211–229.

[12] R. Jin, Y. Xiang, N. Ruan and D. Fuhry, 3-HOP: A high-compression indexing scheme for reachability query, In *Proc.of SIGMOD* (2009), 813–826.

[13] J. Ledlie, P. Gardner and M. Seltzer, Network Coordinates in the Wild, In *Proc.of NSDI*, 2007.

[14] Y. L. Liu, Y. L. Wang and D. J. Guan, An Optimal Fault-Tolerant Routing Algorithm for Double-Loop Networks, *IEEE Trans.on Computers*, 50 (2001), 500–505.

[15] J. A. Nelder and R. Mead, A simplex method for function minimization, *Computer Journal*, 7 (1965), 308–313.

[16] M. Pias, J. Crowcroft, S. Wilbur, T. Harris and S. Bhatti, Lighthouses for scalable distributed location, In *Proc.of IPTPS*, (2003), 278–291.

[17] M. Potamias, F. Bonchi, C. Castillo and A. Gionis, Fast shortest path distance estimation in large networks, In *Proc.of CIKM*, (2009), 867–876.

[18] Z. Qi, Y. Xiao, B. Shao and H. Wang, Fast and scalable analysis of massive social graphs, In *Proc.of VLDB*, (2013), 61–72.

[19] X. L. Ren and L. Y. Lu, Review of ranking nodes in complex networks, *Chinese Science Bulletin*, 13 (2014), 1175–1197.

[20] N. Tse and H. Zhang, Predicting internet network distance with coordinates-based approaches, In *Proc.of INFOCOM*, (2002), 170–179.

[21] F. Wei, TEDI: Efficient shortest path query answering on graphs, In *Proc.of SIGMOD*, (2010), 99–110.

[22] C. K. Wong and D. Coppersmith, A combinatorial problem related to multi-module memory organizations, *Association for Computing Machinery*, 21 (1974), 392–402.

[23] Y. Xiao, W. Wu, J. Pei, W. Wang and Z. He, Efficiently indexing shortest path by exploiting symmetry in graphs, In *Proc.of EDBT*, (2009), 493–504.

[24] X. Zhao, A. Sala, C. Wilson, H. Zheng and B. Y. Zhao, Shortest path estimation for large social graphs, In *Proc.of WOSN*, 2010.

[25] X. Zhao, A. Sala, H. Zheng and B. Y. Zhao, Fast and scalable analysis of massive social graphs, In *Proc.of CORR*, 2011.

[26] X. Zhao, A. Chang, A. Sarma and B. Y. Zhao, On the embeddability of random walk distances, In *Proc.of VLDB*, 6 (2013), 1690–1701.

[27] J. Q. Zhou and X. R. Xu, On infinite families of optimal double-loop networks with non-unit steps, *Ars Combinatoria*, 97A (2010), 81–95.

Received October 2017; revised December 2017.

E-mail address: liuhuihut@163.com
E-mail address: Linzhi@mail.bnu.edu.cn
E-mail address: waqas@mail.bnu.edu.cn