Visualizing Large-scale Network Data Structures

Takeshi Yamada[†]

Abstract

We present a novel approach to visualizing large-scale and complicated data represented by a network in a low-dimensional Euclidean space. It utilizes local connectivity information and produces an efficient visualization that reveals and helps us understand the intrinsic data structure of the original network. A prototype of a three-dimensional Web browser that can represent the Web as a hyperlink network is also described.

1. Introduction

In many scientific and engineering domains, complicated relational data structures are frequently represented by networks or, equivalently, graphs. For example, WWW (World Wide Web) sites are often represented by hyperlink networks, with pages as nodes and hyperlinks between pages as edges; the interactions between genes, proteins, metabolites, and other small molecules in an organism are represented by gene regulatory networks; and the relationships between people and other social entities are characterized by social networks. This is because network representations often provide researchers with important insights for understanding the intrinsic data structure with the help of some mathematical tools such as graph theory. Another simple but important method for studying a network and intuitively understanding its inherent structure is "browsing" over a network layout that is embedded and visualized in a low-dimensional Euclidean space, examining nodes directly one by one, following their connections, and comparing their connectivity with other nodes. Our goal is to develop an efficient visualization algorithm that embeds a network into a low-dimensional Euclidean space in a manner that is suitable for browsing.

Many network visualization methods have been

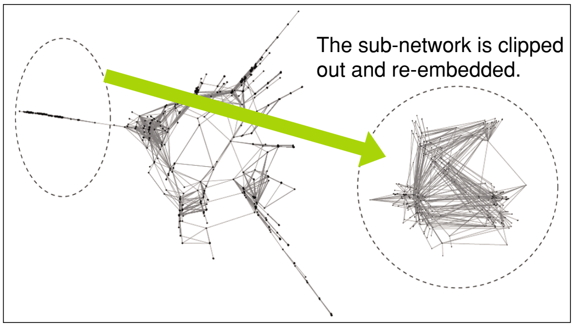


Fig. 1. Visualization result for the NTT network produced by the classical MDS method.

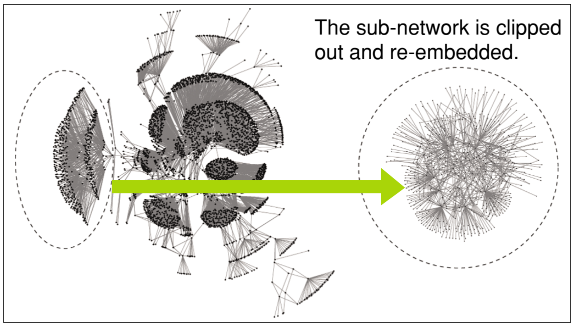


Fig. 2. Visualization result for the NTT network produced by the KK spring method.

[†] NTT Communication Science Laboratories
Soraku-gun, 619-0237 Japan
E-mail: yamada@cslab.kecl.ntt.co.jp

reported in the literature, including simple linear projection methods and more sophisticated spring methods. **Figures 1** and **2** show visualization results for the WWW hyperlink network consisting of all the Web pages located in the NTT domain and including "www.ntt.co.jp" in their URL (universal resource locator), which is referred to as the "NTT network" hereafter. Figure 1 was produced by a linear projection method called classical MDS (multi-dimensional scaling) [1]. We see that many nodes collapse to a single point. This seems to be a limitation of the linear projection methods. Figure 2 was produced by a spring method called the KK (Kamada & Kawai) spring method [2]. It does not suffer from the node collapse problem. However, looking carefully, one can observe that many nodes tend to be arranged densely in semi-circles, which is often characterized as a "dandelion" effect and is not desirable when our aim is browsing. The KK spring method views a network as a graph and first calculates graph-theoretic distances for each pair of nodes. Here, the graph-theoretic distance can be calculated using a shortest path algorithm on a graph. As shown in **Fig. 3**, a virtual spring is then constructed between the nodes where the natural spring length is equal to the graph-theoretic distance. The network layout changes its shape while each virtual spring tries to maintain its natural length as much as possible. As a result, graph-theoretic distances are restored as Euclidian distances. This method is especially effective when the number of nodes is small and a sparse layout is possible. However, it fails for a large-scale network with many nodes and connections.

In both cases, if the sub-network enclosed by the dotted circle on the left of each picture is "clipped out" by removing the nodes and connections outside this circle and is re-visualized as shown on the right in each picture, the embedded network layouts before and after the clipping are observed to become quite different from each other. This observation is expressed by saying that neither the classical MDS method nor the KK spring method has a good "clipping stability" property. While we are browsing a network, we may often want to focus on a small part of the network and study it in detail. A visualization method with poor clipping stability will probably make it very difficult to achieve flexible browsing.

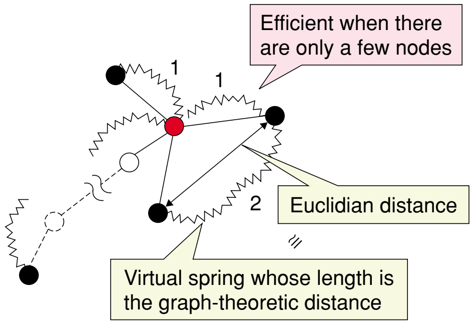

Efficient when there are only a few nodes
1
1
2
Euclidian distance
Virtual spring whose length is the graph-theoretic distance

Fig. 3. The spring method tries to restore graph-theoretic distances as Euclidian distances.

2. Proposed visualization method

NTT Communication Science Laboratories has recently proposed an efficient network visualization method suitable for browsing [3]. This method utilizes only local connectivity information between nodes as a direct criterion. It is based on the following idea: adjacent (i.e., directly connected) nodes should be placed close to each other, and non-adjacent nodes should be placed anywhere as long as they are relatively distant from each other. Here, "relatively distant" means that the distance between non-adjacent nodes is larger than any of the distances between adjacent nodes.

First, a discrete binary similarity measure is defined between nodes such that the similarity between adjacent nodes is 1 and that between non-adjacent nodes is 0. Second, a continuous similarity measure between the embedded nodes is defined such that the Euclidean distance between the nodes is normalized in the range between 0 and 1 by applying some exponential functions. Then a quantity known as the "cross-entropy" of these two similarity measures is calculated. The cross-entropy evaluates the incompatibilities between the two similarity measures. By minimizing the cross-entropy, we can find a network layout embedded in a low-dimensional Euclidean space where the continuous similarity measure approximates the discrete similarity measure as closely as possible, resulting in the two similarity measures being as compatible as possible. For this purpose, a fast and efficient iterative improvement algorithm for minimizing the cross-entropy has been developed. **Figure 4** outlines the proposed method.

Figure 5 shows a visualization result for the same network data as in Figs. 1 and 2, but produced by the proposed visualization method. The nodes in Fig. 5 are laid more uniformly in a radial manner than those in Figs. 1 and 2, giving a space-efficient and well-balanced layout. Moreover, the layouts before and after clipping are fairly similar this time, i.e., the proposed

1. Focus on the local connectivity information instead of using the graph-theoretic distances and consider the two similarity measures:

   - Discrete similarity measure evaluates whether two nodes are adjacent. (i.e., they have a close relationship)
   - Continuous similarity measure evaluates whether their positions are physically close.

2. Define a cross-entropy that evaluates the incompatibilities between the two similarity measures

3. Minimize the cross-entropy by employing the fast iterative improvement algorithm and find the optimal node positions, resulting in a layout such that:

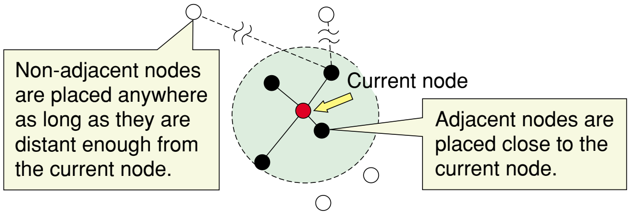

Non-adjacent nodes are placed anywhere as long as they are distant enough from the current node.
Current node
Adjacent nodes are placed close to the current node.

Fig. 4. Outline of the proposed method.

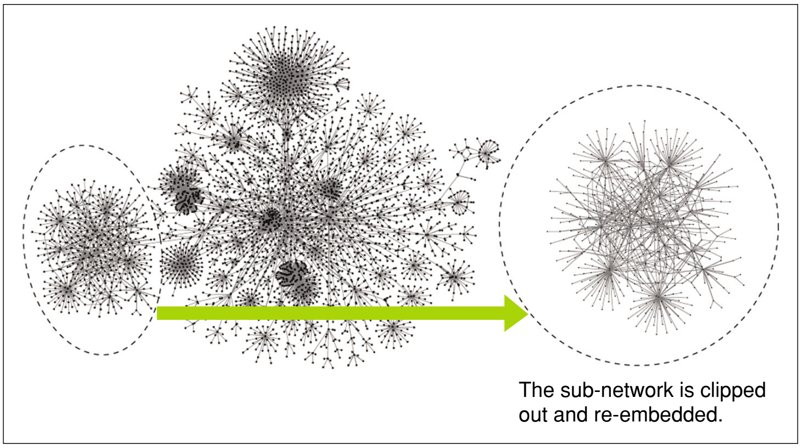


Fig. 5. Visualization result for the NTT network produced by the proposed method.

method has better clipping stability. We believe that clipping stability is an important property for browsing nodes and uncovering new knowledge from the network.

Figure 6 shows a three-dimensional (3D) result for the same data produced by the proposed method. The layout is displayed on a screen using VRML (virtual reality modeling language). You can freely zoom in and out, and the layout can be projected from any viewpoint. Another example is shown in **Fig. 7**, which is human relationship data taken from a Web site [4] that was generated by assembling co-authorship relationships in conference papers that appeared in NIPS (Neural Information Processing Systems) volumes 0 to 12, where an author corresponds to a node and two authors who share at least one joint

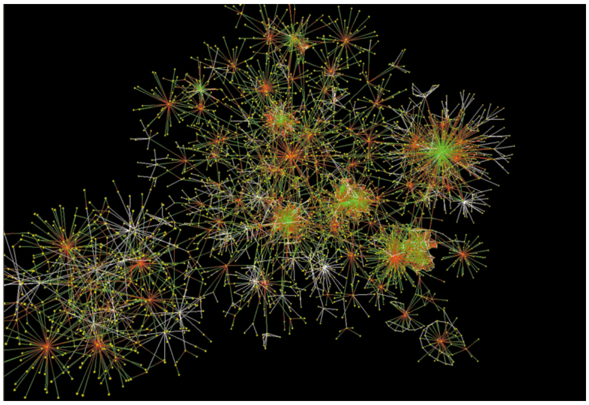


Fig. 6. 3D visualization result for the NTT network produced by the proposed method.

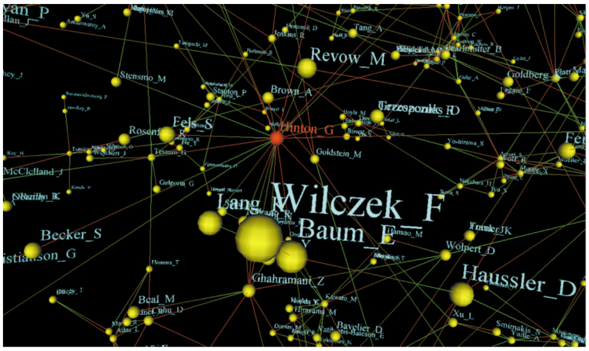


Fig. 7. Visualization result for the NIPS co-authorship network produced by the proposed method.

paper are considered to be directly linked. A portion of the network is displayed in Fig. 7.

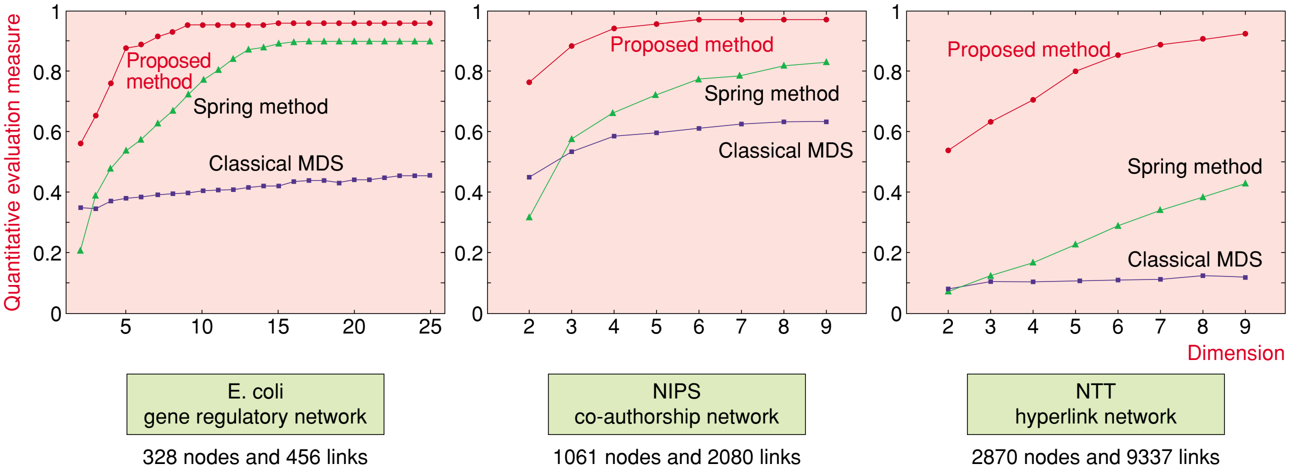
3. Quantitative measure for evaluating visualizations

In the earlier sections, we observed that the proposed visualization method produces an intuitively better network layout. However, those observations were rather subjective. In this section, we attempt to evaluate the embedded network layout in a strictly objective and quantitative fashion. From a visualization viewpoint, embedding into Euclidean spaces with more than three dimensions is meaningless, but quantitative analysis can be applied to embeddings into higher-dimensional spaces.

Experiments were performed to embed different types of network data into relatively low-dimensional spaces by the classical MDS method, KK spring method, and proposed method. Roughly speaking, the proposed evaluation measure is defined to reflect the ratio of “consistent” parts to “inconsistent” parts in the node layout, where a part of the node layout is “inconsistent” when non-adjacent nodes are placed closer than adjacent nodes. **Figure 8** shows the evaluation results. The networks used in the experiment were (a) the gene regulatory network of the bacterium “Escherichia coli” with 328 nodes and 456 links, (b) NIPS co-authorship network with 1061 nodes and 2080 links, and (c) WWW hyperlink network of NTT domain with 2870 nodes and 9337 links. The vertical axis shows the value of the evaluation measure, i.e., the degree of consistency of the embedded network layout; hence, the higher the better. From these results, it is clear that the proposed visualization method always gives the most consistent embeddings, among which the most successful were the results for the WWW hyperlink network of NTT, which is the largest and the most complicated.

4. Application to 3D Web browsers

We wish to use the proposed method to visualize many different types of networks in different areas. As one example, the prototype of our 3D Web browser is shown in **Fig. 9**. The left side of the window is a normal Web browser view and the right side displays the hyperlink network around the currently browsed Web page. A user can move to a new Web page by clicking a node in the network as well as by clicking hyperlink words on the Web page as usual. As the user moves among Web pages, the network grows based on the information gathered so far, and eventu-



E. coli gene regulatory network

328 nodes and 456 links

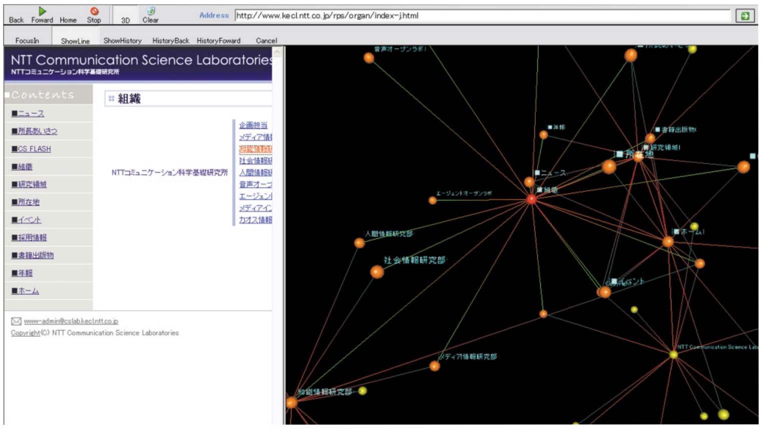
NIPS co-authorship network

1061 nodes and 2080 links

NTT hyperlink network

2870 nodes and 9337 links

Fig. 8. Quantitative evaluations of the visualization results.



Address: http://www.kecl.ntt.co.jp/rps/organ/index-j.html

Fig. 9. Screenshot of the 3D Web browser.

ally the entire hyperlink network structure is revealed.

The system can show the position of the current Web page in the entire hyperlink network. It can also show the history of the visited Web pages as a trajectory and can replay it as needed. It will give us a more realistic and enjoyable Web browsing experience.

The proposed method is also used as a core visualization engine in Multiple Topic Detection by Parametric Mixture Models (PMM), described in the previous article [5].

References

[1] W. S. Torgerson, "Theory and Methods of Scaling," Wiley, New York, 1958.

[2] T. Kamada and S. Kawai, "An algorithm for drawing general undirected graphs," Information Processing Letters, Vol. 32, pp. 7-15, 1989.

[3] T. Yamada, K. Saito, and N. Ueda, "Cross-Entropy Directed Embedding of Network Data," in Proceedings of the Twentieth International Conference on Machine Learning, pp. 832-839, 2003.

[4] S. T. Roweis, "Data for MATLAB hackers: NIPS Conference Papers Vols. 0-12," http://www.cs.toronto.edu/~roweis/data.html, 2002.

[5] K. Saito, "Multiple Topic Detection by Parametric Mixture Models (PMM) — Automatic Web Page Categorization for Browsing," NTT Technical Review, Vol. 3, No. 3, pp. 15-18, 2005.

Takeshi Yamada

Senior Research Scientist, Supervisor, Research Planning Section, NTT Communication Science Laboratories.

He received the B.S. degree in mathematics from the University of Tokyo, Tokyo in 1988 and the Ph.D. degree in informatics from Kyoto University, Kyoto in 2003. In 1988, he joined NTT Electrical Communication Laboratories. He is a member of the Association for Computing Machinery, Institute of Electronics, Information and Communication Engineers of Japan, Scheduling Society of Japan, and Information Processing Society of Japan.