Semantically Modified Diffusion Limited Aggregation for Visualizing Large-Scale Networks

Chaomei Chen and Natasha Lobo
College of Information Science and Technology, Drexel University
3141 Chestnut Street, Philadelphia PA 19104-2875
chaomei.chen@cis.drexel.edu, nrl23@drexel.edu

Abstract

Diffusion-Limited Aggregation (DLA) is a model of fractal growth. Computer models can simulate the fast aggregation of millions of particles. In this paper, we propose a modified version of DLA, called semantically modified DLA (SM-DLA), for visualizing large-scale networks. SM-DLA introduces similarity measures between particles so that instead of attaching to the nearest particle in the aggregation, a new particle is stochastically directed to attach to particles that are similar to it. The results of our initial experiment with a co-citation network using SM-DLA are encouraging, suggesting that the algorithm has the potential as an alternative paradigm for visualizing large-scale networks. Further studies in this direction are recommended.

1. Introduction

Network visualization is a major line of research in information visualization. Just as the Hubble Space Telescope helps astrophysicists to study galaxies in deep space, network visualization techniques enable analysts explore the topology of a network. The most traditional way of visualizing a network is to use node-and-link graphical representations. Graph drawing is an established field that is concerned with how to draw a network algorithmically in compliance with a set of aesthetic criteria [3]. In the field of graph drawing, much of the attention has been given to the efficiency of algorithms and the clarity of end results.

One of the most challenging issues in graph drawing is the scalability of an algorithm in terms of how long it will take for the system to reach equilibrium, to what extent it can minimize the number of links crossing each other, and whether an algorithm can maintain its performance benchmark as the size and/or the density of a network increases. There is still a huge gap, conceptually as well as algorithmically, between structural and dynamic patterns that one would like to access at various levels of granularity.

Diffusion-Limited Aggregation (DLA) is a model of fractal growth. There appears to be some intriguing resemblance between the structure of DLA and the structure of associative network generated in our earlier studies. In this article, we propose a semantically modified DLA (SM-DLA) as a potential alternative method for visualizing large-scale networks. The work presented here is part of our ongoing research in searching for scalable algorithms that can handle large-scale networks. We intend to follow up the present study with more in-depth and more detailed validations and refinements.

2. Graph Drawing Techniques

The most widely known graph drawing techniques include force-directed placement algorithms [14] and spring-embedder algorithms [13, 17]. The primary goal of these types of techniques is to optimize the arrangement of nodes of a network algorithmically, such that in the ultimate geometric model, strongly connected nodes appear close to each other, and weakly connected nodes appear far apart. This optimization is known as the layout process, which is a key topic in the graph drawing community. The strength of the connection between a pair of nodes is typically measured by conceptual similarity, computational relatedness, or conditional probabilities. Examples in this category include Galaxies and SPIRE developed at Pacific Northwest National Laboratories [30]. Multidimensional scaling (MDS) is a closely related topic, significant improvements have been made by latest advances in information visualization [18].

Several public-domain software packages, such as Pajek [1], provide implementations of force-directed graph drawing and spring-embedder algorithms. The strengths of force-directed algorithms include the aesthetic appearance of layout and intuitive good fits between visualization models and underlying network data. The scalability issue is one of the serious drawbacks of these algorithms. Most of

1093-9547/03 $17.00 © 2003 IEEE

these algorithms do not scale well as the size and density of a network increases.

To reduce the complexity of a network visualization task, it is common to prune the original network by link reduction algorithms, or to divide a large network into its smaller components. There are several ways to reduce the number of links in a network: using minimum spanning trees to proximate the original network (MSTs) [19, 21], removing links if their weights are below a threshold [31], or applying network scaling algorithms such as Pathfinder network scaling [22]. Using MSTs can considerably reduce the complexity of network visualization, especially for large-scale networks. Studies using Pathfinder network scaling in the visualization of citation networks include [5, 6, 8-11].

In general, there is a limit for what link reduction can accomplish in resolving the scalability issue. In the divide-and-conquer category, clustering and classification algorithms are the major approaches. Although clustering algorithms can often identify component sub-graphs in a network, there is little that clustering algorithms can do if the network contains giant components and the giant components themselves must be divided. Traditional graph-theoretical algorithms based on connectivity also belong to this category.

Few studies have specifically addressed the challenge of large-scale network visualization. NicheWorks [28] is a notable exception. NicheWorks was designed to visualize networks of tens or hundreds of thousand of nodes. It was tested for analyzing Web sites and detecting international telephone fraud. The strategy taken by NicheWorks is to relax optimization criteria often applied to layout algorithms for smaller networks with the number of nodes ranging from 100 to 1,000.

3. Scientific Networks

Scientific networks are one of the most representative types of complex networks, including scientific collaboration networks, citation networks, research fronts, and knowledge diffusion networks. These networks have been the central subject of several fields of study, notably complex network theory in theoretical physics, scholarly communication and knowledge domain visualization in information science and information visualization.

Scientific networks are large, constantly changing, and practically significant: the citation databases compiled by the Institute for Scientific Information (ISI) alone receive 1 million new publications and 20 million new citations each year; scientific networks evolve in terms of emerging hubs and authorities, competing paradigms, diffusion, and phase transition; and scientific networks have profound connections to a broad spectrum of scientific, technological, social, and economic activities.

Studies of citation networks can be divided into three categories: 1) pioneering studies of citation networks before the 1980s, 2) emergent knowledge domain visualization studies since the late 1990s, and 3) statistical mechanics studies over the last two years. Although there appear to be an increasing number of studies integrating the first two categories, few studies have addressed fundamental issues concerning a potential cross-fertilization between the last two categories.

Derek Price [20] found that the more recent papers tend to be cited about six times more often than earlier papers. The citation rate of a paper then declines. He suggested that scientific literature contains two distinct parts: a classic part and a transient part, and that the two parts have different half-lives. The make-up of these two parts varies from field to field; mathematics, for example, is strongly predominated by the classic part, whereas the transient literature rules in physics. Furthermore, he introduced the notion of research fronts – the collection of highly cited papers that represent the frontiers of science at a particular point of time. Based on an examination of citation patterns of scientific papers, he conjectured that it is possible to identify objectively defined subjects in citation networks. He particularly emphasized the significance of understanding the nature of such moving frontiers in the development of a quantitative method for delineating the topography of current scientific literature.

In the 1970s, Small and Griffith examined issues concerned with identifying specialties by mapping the structure of scientific literatures, especially through analyses of co-citation networks [24]. Small subsequently found rapid changes of focus in collagen research [23]. Documents clustered by their co-citation links can represent leading specialties. The abrupt disappearance and emergence of such document clusters indicate rapid shifts in research focus. By tracing key events through a citation network, Hummon and Doreian [16] successfully re-constructed the most significant citation chain in the development of DNA theory. Their study has great impact on subsequent studies of citation networks in the graph drawing community [2, 4].

The number of studies in knowledge domain visualization is increasing [8-10, 12, 26, 27]. The basic assumption is that the intellectual structure of a scientific community or a subject domain is identifiable through the dynamics of corresponding citation networks. By modeling and visualizing how citation structures change over time, one can explore and better understand how a knowledge domain evolves.

4. Diffusion-Limited Aggregation (DLA)

DLA is a fractal generation model [15, 25]. Diffusion refers to the random motion of a particle in the process of approaching to a cluster of aggregated particles. Individual particles are released one by one to start their diffusion processes. A particle may miss the aggregated cluster and disappear in the void. It may also walk into the cluster. If this happens, the walking particle will be attached to the cluster and the next particle will be launched to repeat a new diffusion process.

Such clusters of particles are called aggregates. Some variations of DLA introduce forces between particles as if they may carry electrical charge. Usually such aggregates are well ordered and similar to the structure of crystals. DLA has a number of practical implications, for instance, in a catalytic process.

According to [29], the shape of a DLA aggregation is in a way controlled by the possibility of particles to reach it. As long there are particles moving around, an aggregate is likely to grow. A diffusing particle is more likely to attach to the outskirt of the aggregation than penetrate deeply to the areas near to the center of the aggregation.

An interesting example of DLA is developed by Chi-Hang Lam[1]. The development of our modified DLA closely references to this particular implementation.

5. DLA for Network Visualization

We are first intrigued by the striking resemblance between the structure of an author co-citation network we generated earlier using Pathfinder network scaling in [7] and the structure of a DLA aggregate.

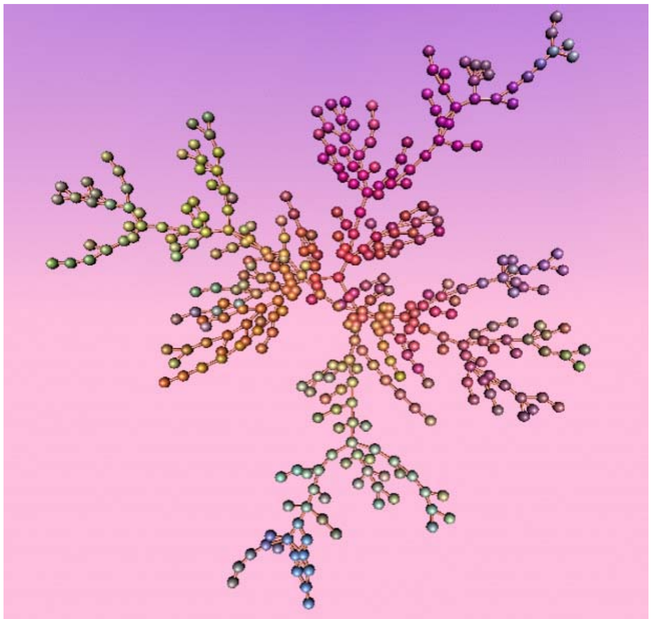

Figure 1. The structure of an author co-citation network [7]. © 1999 IEEE.

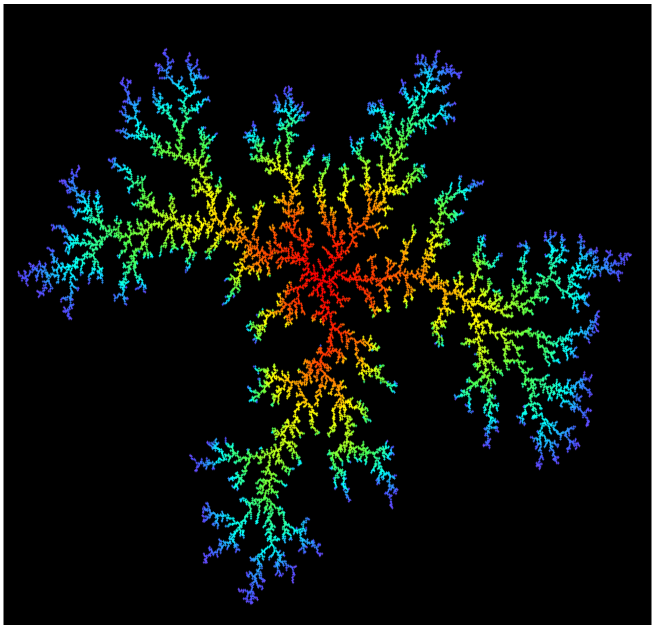

Figure 2. A DLA aggregate of more than 10,000 particles.

Figure 1 shows the Pathfinder representation of the author co-citation network. The most predominant researchers are located in the center of this network, which is a much simplified version of the original network.

Figure 2 is a standard DLA aggregation of more than 10,000 particles. Although we arrived at the two structures from entirely different generating mechanism, the striking resemblance evidently stands out. Both have a sense of a center and a number of branches stretch outwards from the center. An ambitious question is: Is it possible to devise a new type of layout algorithms based on DLA? If we can successfully introduce semantic proximities into the process of DLA, then the aggregation process should still converge and its several appealing features should be inherited by the new approach.

The essence of a DLA algorithm can be summed up by the following pseudo code:

```
launch a new particle $p_0$;
$p_0$ walks randomly until
either
  $p_0$ is close enough to particle p in
  the aggregate:
  attach $p_0$ to p;
or
  $p_0$ is drifted helplessly far away from
  the aggregation:
  re-launch $p_0$;
repeat the procedure;
```

Algorithm 1. A typical DLA algorithm.

[1] http://apricot.polyu.edu.hk/~lam/dla/dla.html

It is clear from the above pseudo code that the attachment is made deterministically in that it is purely based on the distance between the two particles in question. The key idea of our SM-DLA is to introduce some randomness in the attachment so that statistically the new attachment mechanism should be in favor of placing similar particles near to each other.

We refer the modified DLA algorithm as SM-DLA, standing for semantically modified DLA. The new algorithm should meet the following criteria:

1. It should be able to generate an aggregation of a very large network.
2. The probability that two similar particles are placed near to each other should be high.

We undertake two steps in our modification of DLA. First, we impose additional constraint for the random walking particle p_0 to attach to the nearest particle p in the aggregate by introducing randomly generated numbers to simulate the similarity between p_0 and p. The deterministic attachment in the original DLA is now replaced by a probability that is proportional to the semantic similarity. The modified algorithm must still converge and generate an aggregate that is somewhat comparable to the pure DLA. This is a necessary condition for a successful SM-DLA. Second, we make the connection between the randomness in the attachment process and the strength of the similarity between a pair of particles. More detailed descriptions of SM-DLA are given in the following sections.

5.1. Modified DLA – Step One

The first step in modifying the DLA algorithm is outlined in the following pseudo code. The only difference between this version and the original DLA is in the way that an attachment occurs, which is underlined in the pseudo code.

```
launch a new particle $p_0$;
$p_0$ walks randomly until
either
  $p_0$ is close enough to particle p in the
  aggregate:
  generate a random number r;
  if ( r > $threshold_{attach}$ )
    attach $p_0$ to p;
or
  $p_0$ is drifted helplessly far away from
  the aggregation.
  re-launch $p_0$;
repeat the procedure;
```

Algorithm 2. Introducing randomness into the attachment mechanism.

The introduction of the random number in effect raises the threshold for an attachment to take place. The original DLA is equivalent to imposing a threshold of 0, or more precisely any number less than 0. A higher threshold is likely to make the random walking particle keep on walking longer until it hits the right spot. In our experiment, this does not significantly delay the aggregation, which is a good sign for its potential scalability. Figure 3 shows an aggregate produced by the first-step modification of DLA. The attachment threshold is 0.5. This aggregate contains more than 10,000 particles. If it were the size of a network, it would be regarded very large. The color of a particle indicates the time elapsed since the particle arrived in the aggregate. The earlier arrivals are in red, whereas the later ones are in blue. This example shows that the attachments have been made uniformly in all directions.

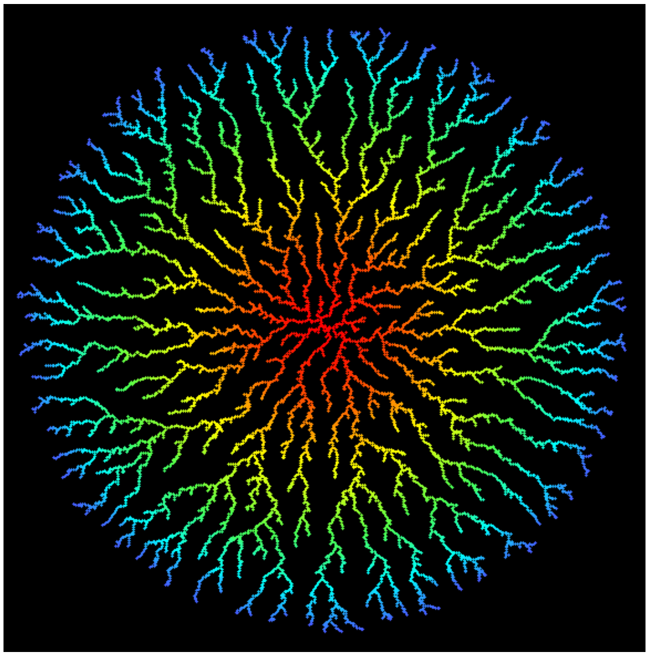

Figure 3. Stochastically modified DLA, containing more than 10,000 particles. The similarity between particles is randomly generated.

5.2. Modified DLA – Step Two

The second step in modifying the DLA algorithm establishes the connection between the attachment probability and the similarity between the random walking particle p_0 and a nearby particle p in the aggregate. The key to the second step is to retrieve the attachment probability directly from the underlying network data.

To ensure that the most similar particles have the best chance to be gathered together, the adjacency list of the network is first sorted in descending order. New particles are launched from the top of the sorted list.

```
sort network edges in descending order of
similarity;
```

```
launch a new particle $p_0$ from the beginning
of the sorted list;
$p_0$ walks randomly until
either
  $p_0$ is close enough to particle p in
  the aggregate:
  if ( sim($p_0$, p) > $threshold_{sim}$ )
    attach $p_0$ to p;
    remove $p_0$ from the sorted list;
  if ( sim($p_0$, p) <= $threshold_{sim}$ )
    generate a random number r
    if ( r > $threshold_{attach}$ )
      attach $p_0$ to p;
      remove $p_0$ from the sorted list;
or
  $p_0$ is drifted helplessly far away from
  the aggregation.
  re-launch $p_0$;
repeat the procedure;
```

Algorithm 3. Semantically modified DLA (SM-DLA).

Figure 4 and Figure 5 show the result of SM-DLA on a real co-citation network of 1,250 articles on botolium toxin research. Each particle represents one article. The similarity between two particles p_i and p_j is defined based on the co-citation strength between articles a_i and a_i as follows:

$sim(p_i, p_j) = cc(a_i, a_j) / sqrt (c(a_i)*c(a_j));$

where cc(a, b) is the co-citation frequency function of articles a and b, and c(a) is the citation function of article a.

The inspection of the result of SM-DLA is encouraging, considering the simple steps in the modification. The shape of the SM-DLA aggregate is essentially similar to but distinct from a typical DLA. The color distribution is an indicator of some features of the SM-DLA aggregate. For example, the color distribution of the aggregate is essentially smooth and continuous. Discontinuous colors indicate the placement of nodes far apart in the sorted pair-wise similarity list. More specifically, a blue particle next to an area of red ones would indicate, probabilistically, that the blue one is likely to a strong similarity to one of the red particles. The conventional DLA does not overlap particles in the aggregate because the nearest particle is all the information the incoming particle needs. In SM-DLA, however, the nearest particle may not be the most similar one. SM-DLA intentionally reduces the probability of the incoming particle landing on such particles. Instead, overlapping particles are allowed so as to give the semantic distance a higher priority than the geometric distance. A possible improvement, not implemented in this version, is to loosen up the aggregate from time to time to allow overlapping particles “fall through the gaps” so that the end result should be more similar to the natural DLA.

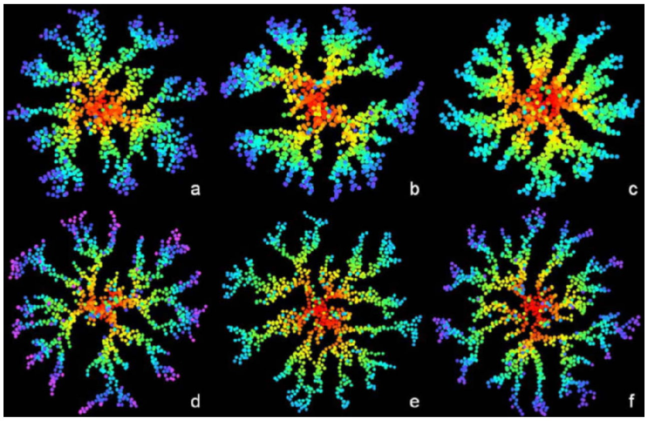

Figure 4. SM-DLA of 1250 articles in botolium toxin research. The similarity between particles is derived from the co-citation strength between corresponding articles. Rows: Threshold$_{attach}$: 9, 8; Columns: Threshold$_{sim}$: 7, 8, and 9.

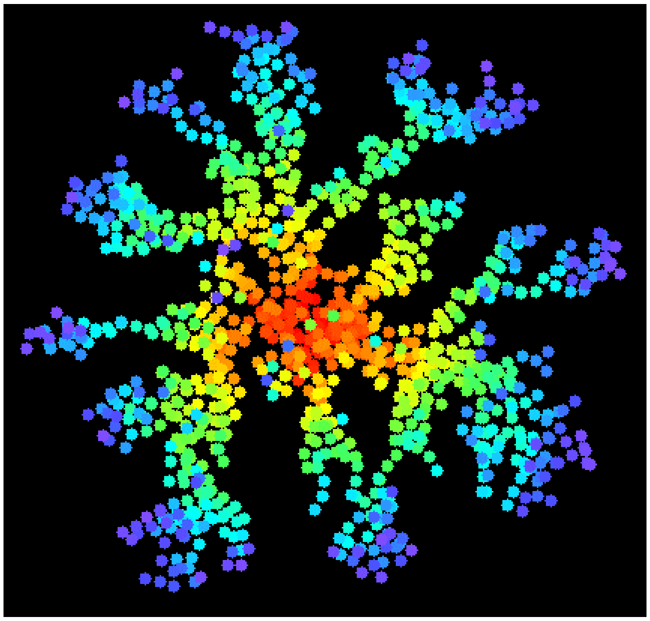

Figure 5. Figure 4a in its original size.

6. Conclusion

In conclusion, we have proposed a semantically modified DLA as a potential alternative for visualizing large networks. In addition to the geometric distance used in the conventional DLA, we have introduced the semantic distance, which takes a higher priority in the aggregation process. We have also introduced the randomness in the attachment mechanism in order to maintain the fundamental aggregation mechanism of the conventional DLA. The initial results appear to suggest that this is a potentially fruitful route to pursue. More thorough experimentations, empirical studies as well as theoretical studies are needed to establish and consolidate the usefulness of a method.

References

[1] V. Batagelj and A. Mrvar, "Pajek - Program for large network analysis," *Connections*, vol. 21, pp. 47-57, 1998.

[2] V. Batagelj and A. Mrvar, "Layouts for GD01 graph-drawing competition," 2001.

[3] G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis, *Graph Drawing: Algorithms for the Visualization of Graphs*: Prentice Hall, 1999.

[4] U. Brandes and T. Willhalm, "Visualization of bibliographic networks with a reshaped landscape metaphor," presented at Proc. 4th Joint Eurographics - IEEE TVCG Symp. Visualization (VisSym '02), 2002.

[5] C. Chen, "Tracking latent domain structures: An integration of Pathfinder and Latent Semantic Analysis," *AI & Society*, vol. 11, pp. 48-62, 1997.

[6] C. Chen, *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. London: Springer-Verlag, 2002.

[7] C. Chen and L. Carr, "Visualizing the evolution of a subject domain: A case study," presented at IEEE Visualization 1999, San Francisco, CA, 1999.

[8] C. Chen, T. Cribbin, R. Macredie, and S. Morar, "Visualizing and tracking the growth of competing paradigms: Two case studies," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 678-689, 2002.

[9] C. Chen and J. Kuljis, "The rising landscape: A visual exploration of superstring revolutions in physics," *Journal of the American Society for Information Science and Technology*, vol. 54, pp. 435-446, 2003.

[10] C. Chen, J. Kuljis, and R. J. Paul, "Visualizing latent domain knowledge," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 31, pp. 518 - 529, 2001.

[11] C. Chen and R. J. Paul, "Visualizing a knowledge domain's intellectual structure," *Computer*, vol. 34, pp. 65-71, 2001.

[12] G. S. Davidson, B. HENDRICKSON, D. K.JOHNSON, C. E. MEYERS, and B. N. WYLIE, "Knowledge mining with VxInsight: Discovery through interaction," *Journal of Intelligent Information Systems*, vol. 11, pp. 259-285, 1998.

[13] P. Eades, "A heuristic for graph drawing," *Congressus Numerantium*, vol. 42, pp. 149-160, 1984.

[14] T. M. J. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software- Practice and Experience*, vol. 21, pp. 1129-1164, 1991.

[15] T. C. Halsey, "Diffusion-limited aggregation: A model for pattern formation," *Physics Today*, vol. 53, pp. 36, 2001.

[16] N. P. Hummon and P. Doreian, "Connectivity in a citation network: The development of DNA theory," *Social Networks*, vol. 11, pp. 39–63, 1989.

[17] T. Kamada and S. Kawai, "An algorithm for drawing general undirected graphs," *Information Processing Letters*, vol. 31, pp. 7-15, 1989.

[18] A. Morrison, G. Ross, and M. Chalmers, "A hybrid layout algorithm for sub-quadratic multidimensional scaling," presented at IEEE Symposium on Information Visualization, Boston, Massachusetts, 2002.

[19] T. Munzner, "H3: Laying out large directed graphs in 3D hyperbolic space," presented at the 1997 IEEE Symposium on Information Visualization, Phoenix, AZ, 1997.

[20] D. D. Price, "Networks of scientific papers," *Science*, vol. 149, pp. 510-515, 1965.

[21] G. G. Robertson, J. D. Mackinlay, and S. K. Card, "Cone trees: Animated 3D visualizations of hierarchical information," presented at CHI '91, New Orleans, LA, 1991.

[22] R. W. Schvaneveldt, "Pathfinder Associative Networks: Studies in Knowledge Organization," in *Ablex Series in Computational Sciences*, D. Partridge, Ed. Norwood, New Jersey: Ablex Publishing Corporations, 1990.

[23] H. G. Small, "A co-citation model of a scientific specialty: A longitudinal study of collagen research," *Scoial Studies of Science*, vol. 7, pp. 139-166, 1977.

[24] H. G. Small and B. C. Griffith, "The structure of scientific literatures I: Identifying and graphing specialties," *Science Studies*, vol. 4, pp. 17-40, 1974.

[25] T. Viczek, *Fractal Growth Phenomena*. Singapore: World Scientific Publishing Co., 1989.

[26] H. D. White and K. W. McCain, "Visualization of literatures," *Annual Review of Information Science and Technology*, vol. 32, pp. 99-168, 1997.

[27] H. D. White and K. W. McCain, "Visualizing a discipline: An author co-citation analysis of information science, 1972-1995," *Journal of the American Society for Information Science*, vol. 49, pp. 327-356, 1998.

[28] G. J. Wills, "NicheWorks: Interactive visualization of very large graphs," *Journal of Computational and Graphical Statistics*, vol. 8, pp. 190-212, 1999.

[29] F.-J. Wirtz, "Diffusion-limited aggregation and its simulation," vol. 2003: Franz-Josef Wirtz, 2003.

[30] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, "Visualizing the non-visual: Spatial analysis and interaction with information from text documents," presented at IEEE Symposium on Information Visualization '95, Atlanta, Georgia, USA, 1995.

[31] M. Zizi and M. Beaudouin-Lafon, "Accessing hyperdocuments through interactive dynamic maps," presented at ECHT '94, Edinburgh, Scotland, 1994.