Generating Functions and Introduction to Evolution Models

Dated : 05.03.07 - 20.03.07

Hema Swetha Koppula

03CS3016 {hema.swetha@gmail.com}

I. Generating Functions: Introduction

A general approach to random graphs with given degree distribution was developed by Newman, Strogatz, and Watts (2001) using a generating function formalism (Wilf, 1990). It turns out that many properties of the network model are exactly solvable in the limit of large network size. The crucial trick for finding the solution is that instead of working directly with the degree distribution p_k, we work with generating functions of the sequence of the degree distribution p_k, which is defined as

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k. \qquad [1]$$

Formalism for calculating various local and global quantities on large unipartite undirected graphs with arbitrary probability distribution of the degrees of their vertices is presented here.

A. Degree Distribution

Considering an uni-partite undirected graph of N vertices, with N large, the generating function $G_0(x)$ for the probability distribution of vertex degrees k is defined as follows:

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k \qquad [2]$$

where p_k is the probability that a randomly chosen vertex on the graph has degree k. The distribution p_k is assumed to be normalized, so that

$$G_0(1) = 1. \qquad [3]$$

The function $G_0(x)$ encapsulates all the information of the degree distribution p_k. Therefore the probability p_k is given by the k^{th} derivative of G_0,

$$p_k = \frac{1}{k!}\frac{d_k G_0}{dx^k}|_{x=0}. \qquad [4]$$

Hence, the function $G_0(x)$ is said to "generate" the probability distribution p_k.

The average degree z of a vertex is nothing but the average over the probability distribution generated by the generating function $G_0(x)$. It is given by

$$z =< k >= \sum_k kp_k = G_0^{'}(1). \qquad [5]$$

Higher moments of the distribution can be calculated from higher derivatives of $G_0(x)$. In general we have

$$< k_n >= \sum_k k^n p_k = [(x\frac{d}{dx})^n G_0(x)]_{x=1} \qquad [6]$$

If a distribution of a property k of an object is generated by a given generating function, then the distribution of the total of k summed over m independent realizations of the object is generated by the m^{th} power of that generating function. This is also referred to as the *Powers* property of the generating functions. Therefore the distribution of the sum the degrees of m randomly chosen vertices is generated by $[G_0(x)]^m$.

B. Distribution of Second Neighbors

Now considering the degree distribution of the vertices that are arrived at by following a random edge. The probability to arrive at a node by a random edge is proportional to the degree of the vertex. Therefore the probability distribution of degree of a vertex is proportional to kp_k. The correctly normalized distribution is generated by

$$\frac{\sum_k kp_k x^k}{\sum_k kp_k} = x\frac{G_o^{'}(x)}{G_0^{'}(1)}. \qquad [7]$$

If we start at a randomly chosen vertex and reach all its k neighbors, then the vertices arrived will have the distribution of remaining outgoing edges generated by this function, less one power of x, to allow for the edge we arrived along. Thus the distribution of outgoing edges is generated by

$$G_1(x) = \frac{G_0^{'}(x)}{G_0^{'}(1)} = \frac{1}{z}G_0^{'}(x), \qquad [8]$$

where z is the average vertex degree.

Thus, the generating function for the probability distribution of the number of *second* neighbors of the original vertex can be written using the ”powers” property of the generating functions as

$$\sum_k p_k[G_1(x)]^k = G_0(G_1(x)). \qquad [9]$$

Similarly, the distribution of third-nearest neighbors is generated by $G_0(G_1(G_1(x)))$. The average number of second neighbors, z_2 is given by

$$z_2 = [\frac{d}{dx}G_0(G_1(x))]_{x=1} = G_0^{'}(1)G_1^{'}(1) = G_0^{''}(1) \qquad [10]$$

C. Examples

C.1 Poisson distributed graphs

In this distribution model the probability $p = z/N$ of the existence of an edge between any two vertices is the same for all vertices. The $G_0(x)$ is given by

$$G_0(x) = \sum_{k=0}^{N} {}^{N}C_k p^k (1-p)^{N-k} x^k \tag*{[11]}$$

$$= (1 - p + px)^N = e^{z(x-1)}, \tag*{[12]}$$

where the last equality applies in the limit $N \longrightarrow \infty$. The average degree of a vertex is $G_0^{`}(1) = z$. In this case

$$G_1(x) = \frac{G_0^{'}(x)}{G_0^{'}(1)} = \frac{ze^{z(x-1)}}{z} = G_0(x). \tag*{[13]}$$

Therefore the distribution of out-going edges at a vertex is the same, regardless of whether we arrived there by choosing a vertex at random, or by following a randomly chosen edge. This property makes the theory of the random graphs simple.

C.2 Exponentially distributed graphs

The exponential distribution of vertex degrees is

$$p_k = (1 - e^{-1/\kappa})e^{-k/\kappa}, \tag*{[14]}$$

where κ is a constant. The generating function for this distribution can obtained as follows:

$$G_0(x) = \sum_k (1 - e^{-1/\kappa})e^{-k/\kappa} x^k \tag*{[15]}$$

$$G_0(x) = (1 - e^{-1/\kappa}) \sum_k^{\infty} e^{-k/\kappa} x^k = \frac{1 - e^{-1/\kappa}}{1 - xe^{-1/\kappa}}, \tag*{[16]}$$

The average degree z is

$$< z >= G_0^{'}(1) = \frac{e^{-1/\kappa}}{1 - e^{-1/\kappa}} \tag*{[17]}$$

and

$$G_1(x) = [\frac{1 - e^{-1/\kappa}}{1 - xe^{-1/\kappa}}]^2 = [G_0(x)]^2 \tag*{[18]}$$

C.3 Power-Law distributed graphs

The distribution is given by $p_k = Ck^{-\tau}e^{-k/\kappa}$ for $k \geq 1$ where $C, \tau and \kappa$ are constants. The exponential cutoff is included as many real-world graphs shoe this cutoff and also it makes the distribution normalizable for all τ. The value of C is fixed as $C = [Li_\tau(e^{-1/\kappa})]^{-1}$ as per the requirement of normalization, where $Li_n(x)$ is the n^{th} polylogarithm of x. $G_0(x)$ can be calculated as

$$G_0(x) = \sum_{k=0}^{\infty} Ck^{-\tau}e^{-k/\kappa}x^k \quad [19]$$

$$= CLi_\tau(xe^{-1/\kappa}) \quad [20]$$

$$= \frac{Li_\tau(xe^{-1/\kappa})}{Li_\tau(e^{-1/\kappa})}. \quad [21]$$

The first The function $G_1(x)$ is given by

$$G_1(x) = \frac{Li_{\tau-1}(xe^{-1/\kappa})}{xLi_{\tau-1}(e^{-1/\kappa})}. \quad [22]$$

The average number of neighbors of a randomly chosen vertex is

$$z = G_0^{'}(1) = \frac{Li_{\tau-1}(e^{-1/\kappa})}{Li_\tau(e^{-1/\kappa})}, \quad [23]$$

and the average number of second neighbors is

$$z_2 = G_0^{''}(1) = \frac{Li_{\tau-2}(e^{-1/\kappa}) - Li_{\tau-1}(e^{-1/\kappa})}{Li_\tau(e^{-1/\kappa})}. \quad [24]$$

D. Component Sizes

Interesting properties like the distribution of the size of the connected components in the graph can be calculated using the generating function approach as discussed here. Let $H_1(x)$ be the generating function for the distribution of the sizes of components which are reached by choosing a random edge and following it to one of its ends. The giant component is excluded from $H_1(x)$. The chances of a component containing a closed loop of edges is N^{-1}, which is negligible in the limit of large N. The distribution of components generated by $H_1(x)$ can be represented as shown in Fig.1; each component is tree-like in structure, consisting of the single site we reach by following our initial edge, plus any number of other tree-like clusters, with the same size distribution, joined to it by single edges. If q_k is the probability that the initial site has k edges coming out of it other than the edge it is arrived through, then $H_1(x)$ can be written as

$$H_1(x) = xq_0 + xq_1H_1(x) + xq_2[H_1(x)]^2 + ... \quad [25]$$

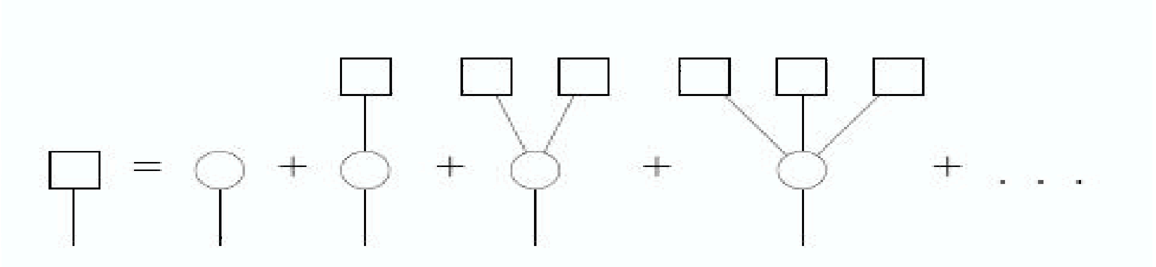

Fig. 1. Schematic representation of the sum rule for the connected component of vertices reached by following a randomly chosen edge.

Since, q_k is the coefficient of x^k in the generating function $G_1(x)$, $H_1(x)$ can be written as

$$H_1(x) = xG_1(H_1(x)). \quad [26]$$

Similarly, the distribution of component size when a node is randomly can be written as

$$H_0(x) = xp_0 + xp_1H_1(x) + xp_2[H_1(x)]^2 + ... \quad [27]$$

$$H_0(x) = xG_0(H_1(x)). \quad [28]$$

The average size of the component a randomly chosen vertex belongs, where there is no giant component, is given by

$$< s >= H_0^{'}(1) = 1 + G_0^{'}(1)H_1^{'}(1). \quad [29]$$

We have

$$H_1^{'}(1) = 1 + G_1^{'}(1)H_1^{'}(1), \quad [30]$$

therefore, we have

$$< s >= 1 + \frac{G_0^{'}(1)}{1 - G_1^{'}(1)} = 1 + \frac{z_1^2}{z_1 - z_2}, \quad [31]$$

where $z_1 = z$ is the average number if neighbors of a vertex and z_2 is the average number of second neighbors.

This equation diverges when $G_1^{'}(1) = 1$. This is the point at which a giant component first appears. Since,

$$G_1^{'}(x) = \frac{\sum_k k(k-1)p_k x^{k-2}}{\sum_k kp_k}, \quad [32]$$

we can also write the condition for phase transition as

$$\sum_k k(k-2)p_k = 0. \quad [33]$$

This sum is a monotonically increasing with the addition of edges to the graph. Therefore, it can be said that the giant component exists if and only if the this sum is positive. From the equation of $< s >$ we also obtain the condition $z_2 > z_1$ for the existence of a giant component.

Till now we have dealt with the case where the giant component hasn't emerged. But the generating function formalism still works when the giant component is present, but according to the definition of $H_0(x)$ which excludes the giant component, $H_0(1)$ will not be unity. $H_0(1)$ in this case would be equal to $1 - S$, where S is the fraction of the graph occupied by the giant component.

We can write

$$S = 1 - G_0(u) \qquad [34]$$

, since $H_0(x) = G_0(H_1(x))$, where $u \equiv H_1(1)$ is the smallest non-negative real solution of

$$u = G_1(u). \qquad [35]$$

The general expression for the average component size excluding the giant component is given as follows,

$$< s >= \frac{H_0^{'}(1)}{H_0(1)} \qquad [36]$$

$$= \frac{1}{H_0(1)}[G_0(H_1(1)) = \frac{G_0^{'}(H_1(1))G_1(H_1(1))}{1 - G_1^{'}(H_1(1))}] \qquad [37]$$

$$= 1 + \frac{zu^2}{[1 - S][1 - G_1^{'}(u)]}. \qquad [38]$$

This is the same as the equation derived for $< s >$ in absence of giant components in which case $S = 0$, and $u = 1$.

E. Number of Neighbors and Average Path Length

The number of neighbors which are at a distance of m steps from a randomly chosen vertex can be expressed as follows:

$$G(x)^{(m)} = G_0(G_1(...G_1(x)...)) \qquad [39]$$

Therefore $G^{(m)}(x)$ can be defined as the generating function for the m^{th} neighbor.

$$G(x)^{(m)} = \begin{cases} G_0(x), & \text{for m = 1;} \\ G^{m-1}(G_1(x)), & \text{for } m \geq 2. \end{cases} \qquad [40]$$

The average number of m^{th} neighbors, z_m is

$$z_m = \frac{dG^m}{dx}|_{x=1} = G_1^{\prime}(1)G^{(m-1)^{\prime}}(1) = G_1^{\prime}(1)z_{m-1} \qquad [41]$$

Along with the initial condition $z_1 = z = G_0'(1)$. Therefore,

$$z_m = [G_1'(1)]^{m-1} G_0'(1) = [\frac{z_2}{z_1}]^{m-1} z_1. \tag{42}$$

Now we can make an estimate the length of the shortest path between two randomly chosen vertices on the graph, l. This length is reached approximately when the total number of neighbors of a vertex out to that distance is equal to the number of vertices on the graph, i.e., when

$$1 + \sum_{m=1}^{l} z_m = N. \tag{43}$$

Therefore,

$$z_1 + z_1.\frac{z_2}{z_1} + z_1.(\frac{z_2}{z_1})^2 + ... + z_1.(\frac{z_2}{z_1})^{l-1} = N - 1 \tag{44}$$

$$z_1.\frac{1 - (\frac{z_2}{z_1})^l}{1 - \frac{z_2}{z_1}} = N - 1 \tag{45}$$

On rearranging,

$$(\frac{z_2}{z_1})^l = 1 - \frac{N-1}{z_1^2}.(z_1 - z_2) \tag{46}$$

On taking log both sides and rearranging, we have

$$l = \frac{\log[(N-1)(z_2 - z_1) + z_1^2] - \log z_1^2}{\log \frac{z_2}{z_1}} \tag{47}$$

In the common case where $N \gg z_1$ and $z_2 \gg z_1$, this results

$$l = \frac{log(N/z_1)}{log(z_2/z_1)} + 1 \tag{48}$$

This method assumes that all vertices are reachable from a randomly chosen starting vertex, which is not true in general unless there is a giant component which fills the entire graph. And also the conditions used to derive are only an approximation, hence this result is only an approximation. Even with such shortcomings it has a number of remarkable features: i)average vertex-vertex distance for all random graphs scale logarithmically with the size of N, regardless of the degree distribution, ii) the average distance, which is a global property, can be calculated using only the knowledge of the average number of first- and second-nearest neighbors, which are local properties, and iii) two random graphs with completely different distribution of vertex degrees, but the same values of z_1 and z_2, will have the same average distances.

II. Models of Network Growth: Introduction

The class of models discussed here aim at explaining the network properties by examining the growth process of the network by gradual addition of vertices and edges as might be taking place on the real networks. It is the growth process which lead to the characteristic structural features of a network.

A. Price's Model:

This model talks about directed networks like the citation network. Derek de Solla Price studied the *citation networks* of scientific papers and found that both in-degrees and out-degrees have power-law distributions. His idea was that a paper which has more number of citations will have more probability to get cited in a newly written paper. He called this the *cumulative advantage*, which is usually known as the *preferential attachment.*

B. The Model of Barabasi and Albert:

The Barabsi-Albert model incorporates two important general concepts: growth and preferential attachment. Both growth and preferential attachment exist widely in real networks.

Growth means that the number of nodes in the network increases over time. *Preferential attachment* means that the more connected a node is, the more likely it is to receive new links. Nodes with higher degree have stronger ability to grab links added to the network. Intuitively, the preferential attachment can be understood if we think in terms of social networks connecting people. Here a link from A to B means that person A "knows" or "is acquainted with" person B. Heavily linked nodes represent well-known people with lots of relations. When a newcomer enters the community, s/he is more likely to become acquainted with one of those more visible people rather than with a relative unknown.

Preferential attachment is an example of a positive feedback cycle where initially random variations (one node initially having more links or having started accumulating links earlier than another) are automatically reinforced, thus greatly magnifying differences. This is also sometimes called the Matthew effect, "the rich get richer".

The algorithm used in the BA model goes as follows.

1. Growth: Starting with a small number m_0 of connected nodes, at every time step, we add a new node with $m(< m_0)$ edges that link the new node to m different nodes already present in the network.

2. Preferential attachment: When choosing the nodes to which the new node connects, we assume that the probability P that a new node will be connected to node i depends on the degree k_i of node i, such

that

$$P \sim \frac{k_i}{\sum_i k_i} \qquad [49]$$

This model differs from the Price's Model in considering undirected edges. By considering undirected edges, the problem Price's model of how a paper gets its first citation is solved. Each vertex in the graph has an initial degree of m, and hence has a non-zero probability of receiving new links. There are also disadvantages for considering undirected edges as the citation networks and the Web are directed in nature, and hence the model is missing out a crucial feature of these networks.

Example 1: Taking a m_0 as 4 and m as 3, the step by step addition of nodes according to the BA model is shown in the figures Fig2, Fig3, Fig4, Fig5, and Fig6. The probability with which each node might get a new edge is given next to each node.

It can be clearly seen that the nodes with higher degree tend to get new edges at each step than the ones with lower degree. Therefore, "the rich gets richer". Due to this property hubs will be formed. In this example, nodes 3 and 4 form hubs.

Using the Master Equation Method, the Barabasi and Albert model can be solved analytically as follows:

The probability that a new edge attaches to a vertex of degree k is:

$$\frac{kp_k}{\sum_k kp_k} = \frac{kp_k}{2m} \qquad [50]$$

Where $\sum_k kp_k$ is equal to the mean degree of the network, which is 2m. The mean number of vertices of degree k that gain an edge when a single new vertex with m edges is added is $m * kp_k/2m = \frac{1}{2}kp_k$, independent of m. Therefore the number of vertices with degree k will decrease by this amount as these vertices get an extra edge and no longer of degree k. There is also an influx from vertices previously of degree k-1 and have acquired a new edge. The boundary case occurs when k = m, which have an influx of exactly 1. If the value of p_k when the graph has n vertices is denoted by $p_{k,n}$, then the net change in np_k per vertex added is

$$(n+1)p_{k,n+1} - np_{k,n} = \frac{1}{2}(k-1)p_{k-1,n} - \frac{1}{2}kp_{k,n}, \qquad [51]$$

for $k > m$, or

$$(n+1)p_{m,n+1} - np_{m,n} = 1 - \frac{1}{2}kp_{m,n}, \qquad [52]$$

for k = m. Now, looking for stable solutions $p_{k,n+1} = p_{k,n} = p_n$, we have

$$p_k = \begin{cases} \frac{1}{2}(k-1)p_{k-1} - \frac{1}{2}kp_k, & \text{for } k > m; \\ 1 - \frac{1}{2}mp_m, & \text{k = m.} \end{cases} \qquad [53]$$

Rearranging for p_k, we get $p_m = \frac{2}{m+2}$ and $p_k = p_{k-1}(k-1)/(k+2)$, or

$$p_k = \frac{(k-1)(k-2)...m}{(k+2)(k+1)...(m+3)} p_m = \frac{2m(m+1)}{(k+2)(k+1)k} \qquad [54]$$

In the limit of large k this gives a power law degree distribution $p_k \sim k^{-3}$.

Therefore the results indicate that this network evolves into a scale-invariant state with the probability that a node has k edges following a power law with an exponent $\alpha = 3$. The scaling exponent is independent of m, the only parameter in the model.

III. References

1. "Random graphs with arbitrary degree distribution and their applications" - M. E. J. Newman, S. H. Strogatz & D. J. Watts.
2. "The structure and function of complex networks" - M. Molloy & B. Reed.
3. "Random graph models of social networks" - M. E. J. Newman , D. J. Watts & S. H. Strogatz.
4. "BA Model on Wikipedia" - $http : //en.wikipedia.org/wiki/Preferential_attachment.$
5. "Statistical mechanics of complex networks" - Reka Albert & Albert-Laszlo Barabasi.

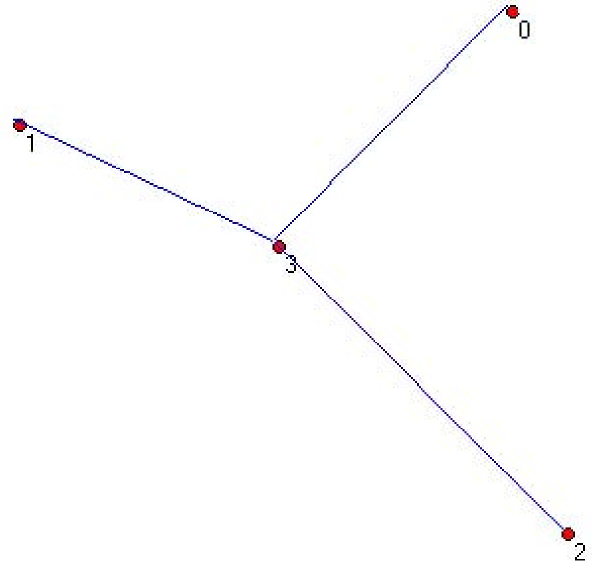

Fig. 2. Step1: Initial nodes : 0, 1, 2, 3.

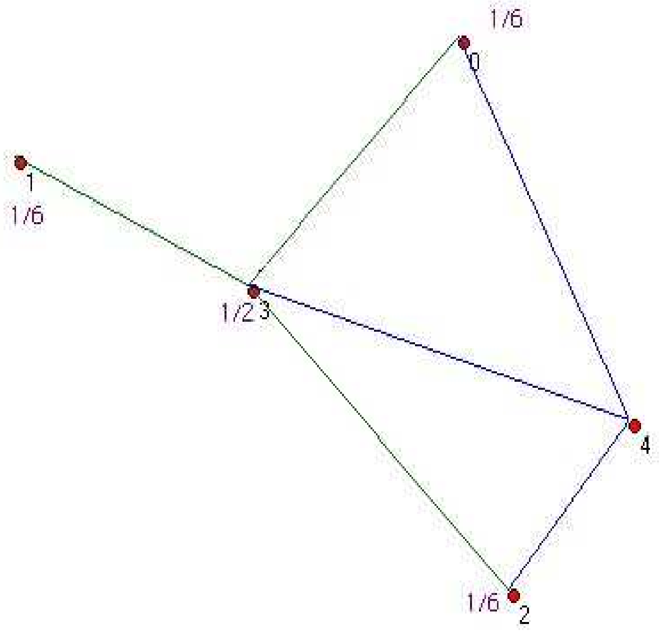

Fig. 3. Step2: Node 4 is added.

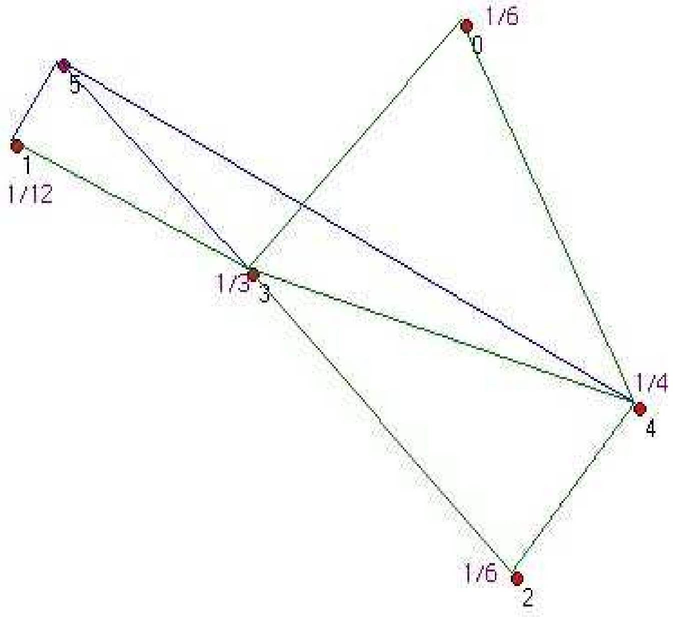

Fig. 4. Step3: Node 5 is added.

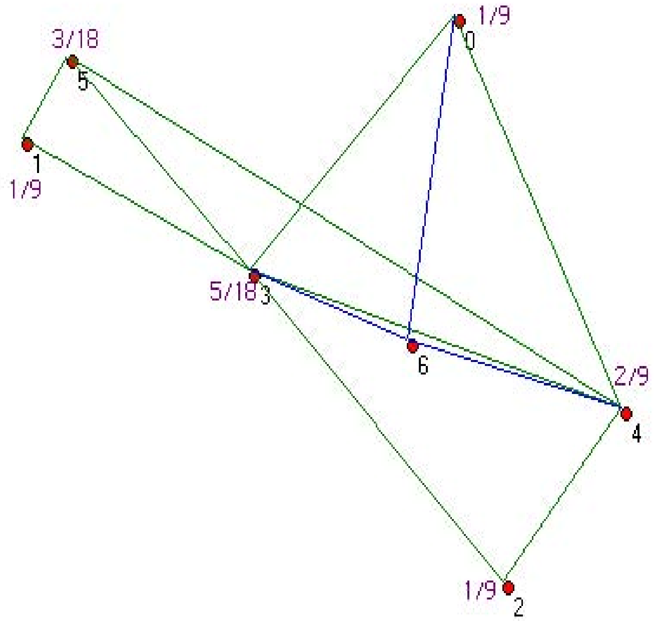

Fig. 5. Step4: Node 6 is added.

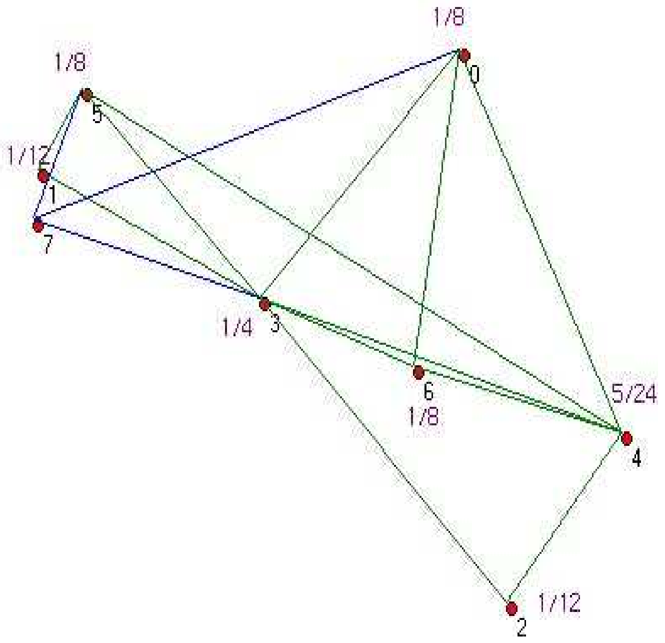

Fig. 6. Step5: Node 7 is added.