Social Network Graph Analysis Using SQL Report

Inserting the Data

- The tables facebook_5037 and twitter_5037 was created with the columns ‘source_id’ and ‘destination_id’ using the following sql queries

```
✓ CREATE TABLE facebook_5037 (
   source_id int NOT NULL,
   destination_id int NOT_NULL
  );

✓  CREATE TABLE twitter_5037(
    source_id int NOT NULL,
    destination_id int NOT_NULL
    );
```

- The data that has to be inserted in the tables facebook_5037 and twitter_5037 is extracted from the network files facebook_network.txt and twitter_network.txt, which are available on /data/teaching/uxac007/ on bigdata.

  The command to insert data to the tables facebook_5037 and twitter_5037 from the files is as follows:

```
✓ \COPY facebook_5037 FROM /data/teaching/uxac007/facebook_network.txt delimiter ’ ’

✓ \COPY twitter_5037 FROM /data/teaching/uxac007/twitter_network.txt  delimiter ’ ’
```

- Schema of facebook_5037 is

```
 source_id | destination_id
-------------+----------------
```

  Schema of twitter_5037 is

```
 source_id | destination_id
------------+----------------
```

(1) Counting Triangles

The undirected graph G has a total number of nodes as ‘n’ and total number of edges as ‘m’. There can be nC_3 or $(^n{}_3)$ possible triangles in G. The probability(p) of any two nodes being connected by an edge is m / (nC_2) or m / $(^n{}_2)$. Triangle has three edges, hence the multiplication rule produces $(m / (nC_2))^3$ or $(m / (^n{}_2))^3$. Hence, the expected number of triangles in G is (nC_3) x $(m / (nC_2))^3$ or $(^n{}_3)$ x $(m / (^n{}_2))^3$. So, the values of total number of nodes (n) and the total number of edges (m) have to be calculated as follows:

Total number of nodes (n)

The total number of nodes (n) in the facebook graph can be calculated using the following query and we get the value of n as 4039.

Query:

```
with a as(select source_id from facebook_5037 union select destination_id from
facebook_5037)select count(distinct source_id) as n from a;
```

Output:

```
   n
 ------
 4039
(1 row)
```

Total number of edges (m)

The total number of edges (m) between the nodes in the facebook graph can be calculated using the following query and we get the value of m as 88234.

Query:

```
select count(source_id) as m from facebook_5037;
```

Output:

```
m
-------
88234
(1 row)
```

(1.1) Expected Number of Triangles in Random Graph (n and m values as Facebook Graph)

The expected number of triangles in random graph is calculated as (nC_3) x $(m / (nC_2))^{3}$ or $(^{n}\ _{3})$ x $(m / (^{n}\ _{2}))^{3}$. The values of n and m in the random graph are the same as for the facebook graph, so the value of n is 4039 and the value of m is 88234.

So, we can compute the expected number of triangles manually using the given formula to get the value as 13900.3282 . The value of expected number of triangles can also be calculated using the following query:

Query:

```
with a as(select count(distinct source_id) as n,(select count(source_id) as m from facebook_5037)from(select source_id from facebook_5037 union select destination_id from facebook_5037) as foo1)select d*(b/c) as expected_num_of_triangles from(select (n*(n-1)*(n-2))/6 as d, m^3 as b, ((n*(n-1))/2)^3 as c from a)as foo;
```

Output:

```
expected_num_of_triangles
    --------------------------
     13900.3281973242
              (1 row)
```

(1.2) Actual Number of Triangles in Facebook Graph

The triangles in the facebook graph are formed when three nodes are connected to each other. So, the actual number of triangles are formed in the facebook graph is found using the following query:

Query:

```
select count(*) as actual_num_of_traingles from facebook_5037 t1,facebook_5037 t2,facebook_5037 t3 where t1.destination_id=t2.source_id and t2.destination_id=t3.destination_id and t1.source_id=t3.source_id;
```

Output:

```
actual_num_of_triangles
    ------------------------
            1612010
            (1 row)
```

(1.3) Comparison and Interpretation

The expected number of triangles from a random graph which has the same number of nodes (n) and same number of edges (m) is 13900.3281973242 and the actual number of triangles in facebook graph is 1612010. This shows that the expected number of triangles of random graph is much less than the actual number of triangles in facebook graph.

If A and B are friends, B and C are friends, then A and C are likely to be friends. This likeliness of a triangle being formed in the random graph is calculated in the expected number of triangles in the random graph, which has the same number of nodes and edges as facebook graph.

Even though the random graph has the same number of nodes and edges as the facebook graph, the likeliness of an edge being formed or the probability of an edge summing up together to give the expected number of triangles as 13900.3282 whereas the facebook graph formed from the dataset forms 1612010 triangles actually. So, the expected number of triangles of a random graph similar to facebook graph, that is, random graph with same n and m as facebook graph is less than the actual number of triangles in facebook graph.

(2) Degree Visualisations

2.1 Degree Distribution in Facebook Graph

(2.1.1) Binomial Vs Poisson Distribution of Random graph (n and m as Facebook graph)

The probability of an edge being formed between two nodes in the facebook graph is p : m / (nC_2) or m / ($^{n}{}_{2}$). Let X be any node among the n number of nodes, and let it have k number of incident nodes among the other (n-1) nodes. So, the binomial degree distribution is given by
$P(k)={}^{(n-1)}C_k\ p^k\ q^{(n-1-k)}$
And the poisson distribution is given by
$P(k)= e^{-lambda}\ (lambda)^k/k!$
Where lambda= np

To find the binomial and the poisson distribution of a random graph, we need the values of n, m and k. Since the random graph has the same value of total number of nodes (n) and total number of edges (m) as facebook graph, n is 4039 and m is 88234. Now the degree of each node has to be calculated, which is calculated using the following query:

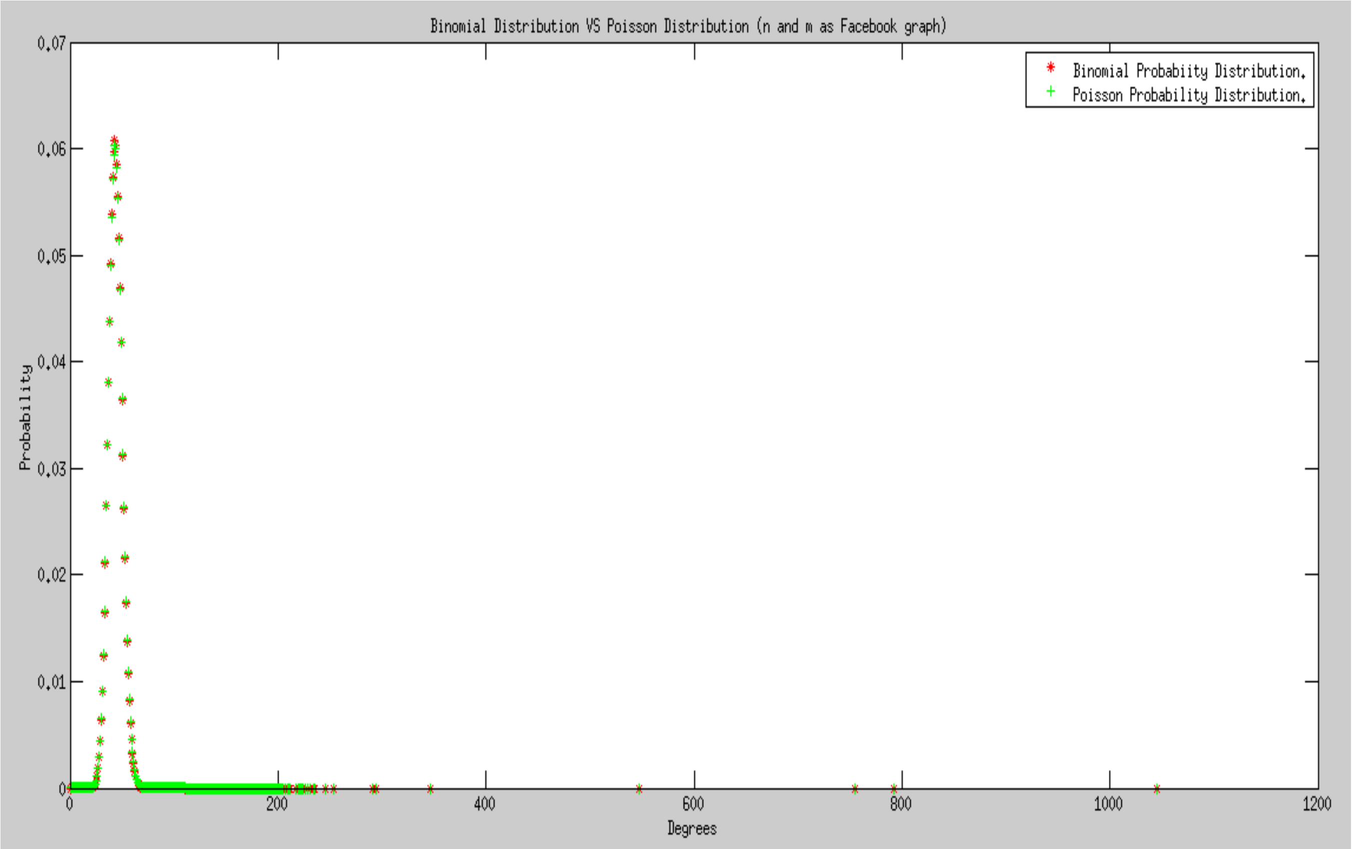
Query:

```
select source_id,sum(degree) as degree from (select source_id,count(source_id) as degree from facebook_5037 group by(source_id) union all select destination_id,count(destination_id) as degree from facebook_5037 group by(destination_id) order by(source_id))as foo group by(source_id);
```

This query output, which is the degree for all the nodes in the random graph must be stored in a file so that plotting can be done for the degrees and the binomial and poisson distributions. The following command is used to store the output of the query in a file:

Storing results of query in a file:

```
\COPY (select source_id,sum(degree) as degree from (select source_id,count(source_id) as degree from facebook_5037 group by(source_id) union all select destination_id,count(destination_id) as degree from facebook_5037 group by(destination_id) order by(source_id))as foo group by(source_id)) TO '/rmt/csfiles/pgrads/mbva620/Ass3-2.csv'WITH CSV;
```



The above graph shows the plot between the degrees in x-axis and the probability (P(k)), which can be Binomial Probability Distribution (or) Poisson Probability Distribution in the y-axis. The red (*) denotes the Binomial Distribution and the green (+) denotes the Poisson Distribution. The comparison between the distribution is done as follows:

Comparison of Results

- The graph clearly shows that the binomial and the poisson distribution looks similar and overlaps at many instances. This is because binomial and poisson distribution are the same but poisson distribution varies with large values of nodes(n). Since the values of n are the same in both the distributions, both binomial and poisson distributions mostly overlap with each other but there are a few differences in the distribution.

- We can observe that binomial distribution has the probability '0' when its degree is 0 and for a few degrees between 200 and 400, in 550, in 760 and 795, and finally at the degree 1050 but poisson distribution has the probability 0 at many degrees such as from 0 to 40, from 90 to more than 200 and the values of degrees where the binomial distribution also had probability 0 such as in 550, 760,795 and 1050.

- Then almost all the values of binomial distribution overlaps with poisson distribution except a few point differences between the two distributions. For instance, at degree 50, the value of binomial distribution is around 0.062 and it is 0.061 for poisson distribution, which is the highest probability of both binomial and poisson distribution. Similarly, there is only a small difference between the binomial and poisson distribution for each degree but it also seems that the binomial distribution exceeds the poisson distribution in most cases by a very small value. Also if the value of n (total number of nodes) is increased, the poisson distribution, this is because as n increases, p value decreases and the computation value of ‘lambda’, which is used for the calculation of P(k) in poisson distribution decreases further. In the above graph, the p value was 0.0108, which gave the high peaks in poisson distribution and the split in binomial distribution.

(2.1.3) Degree Distribution of Actual Facebook Graph

The degree distribution of actual facebook graph can be found by taking the degrees that are present for each node in the facebbok graph. Then, taking into account all the degrees in total, the number of occurrences of each degree is taken as the count. This count is divided by the total number of nodes to find the proportion of the degree occurrence in the facebook graph. This proportion of each degree can be calculated as follows:

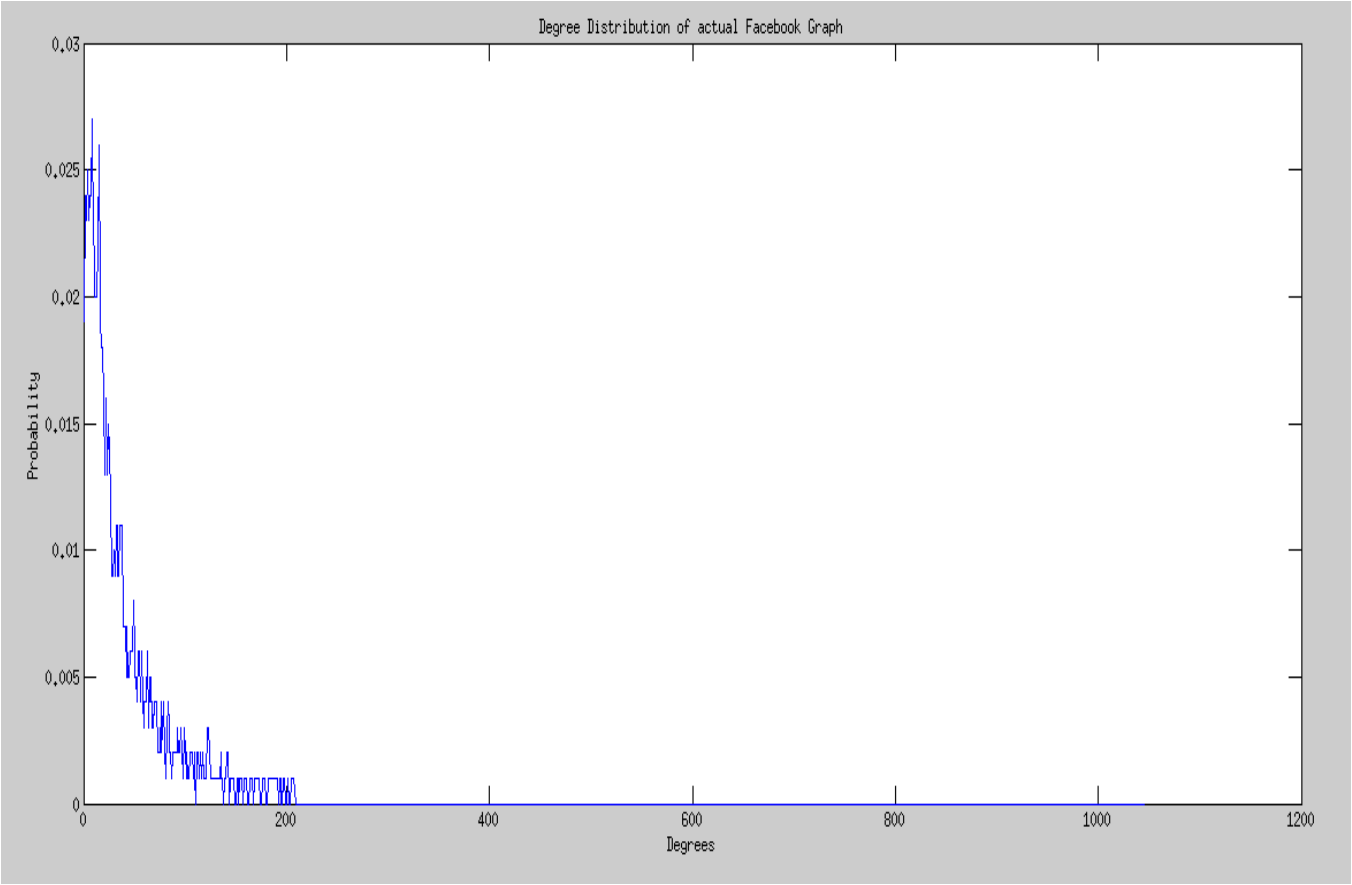
Query:

```
with a as(select source_id,sum(degree) as degree from (select source_id,count(source_id) as
degree from facebook_5037 group by(source_id) union all select
destination_id,count(destination_id) as degree from facebook_5037 group
by(destination_id) order by(source_id))as foo group by(source_id))select
degree,round((count(source_id)/(with b as(select source_id from facebook_5037 union
select destination_id from facebook_5037)select count(distinct source_id) as n from
b)::numeric),3)from a group by(degree) order by(degree);
```

This query outputs the degree and the proportion of that degree occurrence in the facebook graph. This output has to be stored in a file to plot the degree distribution of facebook graph. The following query is used to store the output in a file:

Storing results of query in a file:

```
\COPY (with a as(select source_id,sum(degree) as degree from (select
source_id,count(source_id) as degree from facebook_5037 group by(source_id) union all
select destination_id,count(destination_id) as degree from facebook_5037 group
by(destination_id) order by(source_id))as foo group by(source_id))select
degree,round((count(source_id)/(with b as(select source_id from facebook_5037 union
select destination_id from facebook_5037)select count(distinct source_id) as n from
b)::numeric),3)from a group by(degree) order by(degree)) TO
'/rmt/csfiles/pgrads/mbva620/Ass3.csv'WITH CSV;
```



The degree distribution of actual facebook graph is formed by plotting the degrees in the x-axis and the proportion of the degrees' occurrence, that is, (number of times the degree has occurred across all the nodes in the graph)/ (total number of nodes(n)).

(2.1.4) Interpretations

- The degrees are distributed in the actual facebook graph is similar to the power law graph, that is, P(k) ~ x^{-alpha}, where alpha < 1, with a long tail. The power law, according to statistics, is a functional relationship between two quantities, where one quantity varies as a power of another. So, the degree distribution of facebook graph looks similar to the linear scale of the Power law distribution and so if we plot the degree distribution of facebook graph in log-log scale, then it is almost a diagonal line connecting the x and the y-axis.

- The degree distribution of facebook graph does not exactly match the poisson distribution, whereas it looks similar to a graph plotted by power law distribution. When power law is plotted, the initial values y-axis values are very high with small values of the term in x-axis and then the value in y-axis decreases gradually with an increase in x-axis values. Hence, the above graph on degree distribution of facebook graph shows that the shape of this degree distribution matches most with the power law distribution.

- Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The above graph shows a positive skew or right skewed, that is, there is a long tail on the right side of the graph. The right tail is longer; the mass of the distribution is concentrated on the left of the figure. So, larger share of population rests within its tail, that is, many degrees are covered (from 200 to around 1050) by the long tail on the right side of the graph. We can also observe that these degrees on the long tail have the probability 0 in the degree distribution. Since the degree distribution is positively skewed, the mean is greater than the median reflecting the fact that the mean is sensitive to each degree in the distribution and is subject to large shifts when the sample is small and contains extreme degrees.

- There are a few hubs in the facebook graph, that is, the nodes which are exceptionally well connected to the other nodes in the graph. The above degree distribution of facebook graph is a plot between the degrees and the proportion of nodes of each degree (count of each degree / total number of nodes). So, it can be observed from the above graph that the many nodes had the degree around 0 to 200 and then it declined. The node which has the highest degree is the hub as it will be well connected among various nodes in the graph, such a node is '107' which has the highest degree of 1045. The other hubs or well- connected nodes are '1684' with the degree of 792 and '1912' with the degree of 755.

2.2 Degree Distribution in Twitter Graph

(2.1.1) Binomial Vs Poisson Distribution of Random graph (n and m as Twitter graph)

The probability of an edge being formed between two nodes in the twitter graph is p :
m / (nC_2) or m / ($^{n}{}_{2}$). Let X be any node among the n number of nodes, and let it have k as indegree among the other (n-1) nodes. So, the binomial degree distribution is given by
$P(k)={}^{(n-1)}C_k\ p^k\ q^{(n-1-k)}$
Similarly, the same formula is used for k as outdegree among the other (n-1) nodes.

And the poisson distribution is given by
$P(k)= e^{-lambda}\ (lambda)^k/k!$
Where lambda= np.
Similarly, the same formula is used for k as outdegree.

To find the binomial and the poisson distribution of a random graph, we need the values of n, m and k (as indegree as well as outdegree). Since the random graph has the same value of total number of nodes (n) and total number of edges (m) as twitter graph, n is 81306 for indegree, n is 70097 for outdegree and m is 2420766. The values of n and m are calculated and then the indegree and outdegree are calculated using the following queries:

Query:

Total number of nodes (n):

```
with a as(select source_id from twitter_5037 union select destination_id from twitter_5037)select count(distinct source_id) as n from a;
```

Output:

```
   n
-------
 81306
(1 row)
```

Query:

Total number of edges (m):

```
select count(source_id) as m from twitter_5037 where source_id != destination_id;
```

Output:

```
    m
---------
 2420744
(1 row)
```

(2.2.1) Degree Distribution of random graph with same number of nodes (n) and edges (m) as Twitter graph

Query:

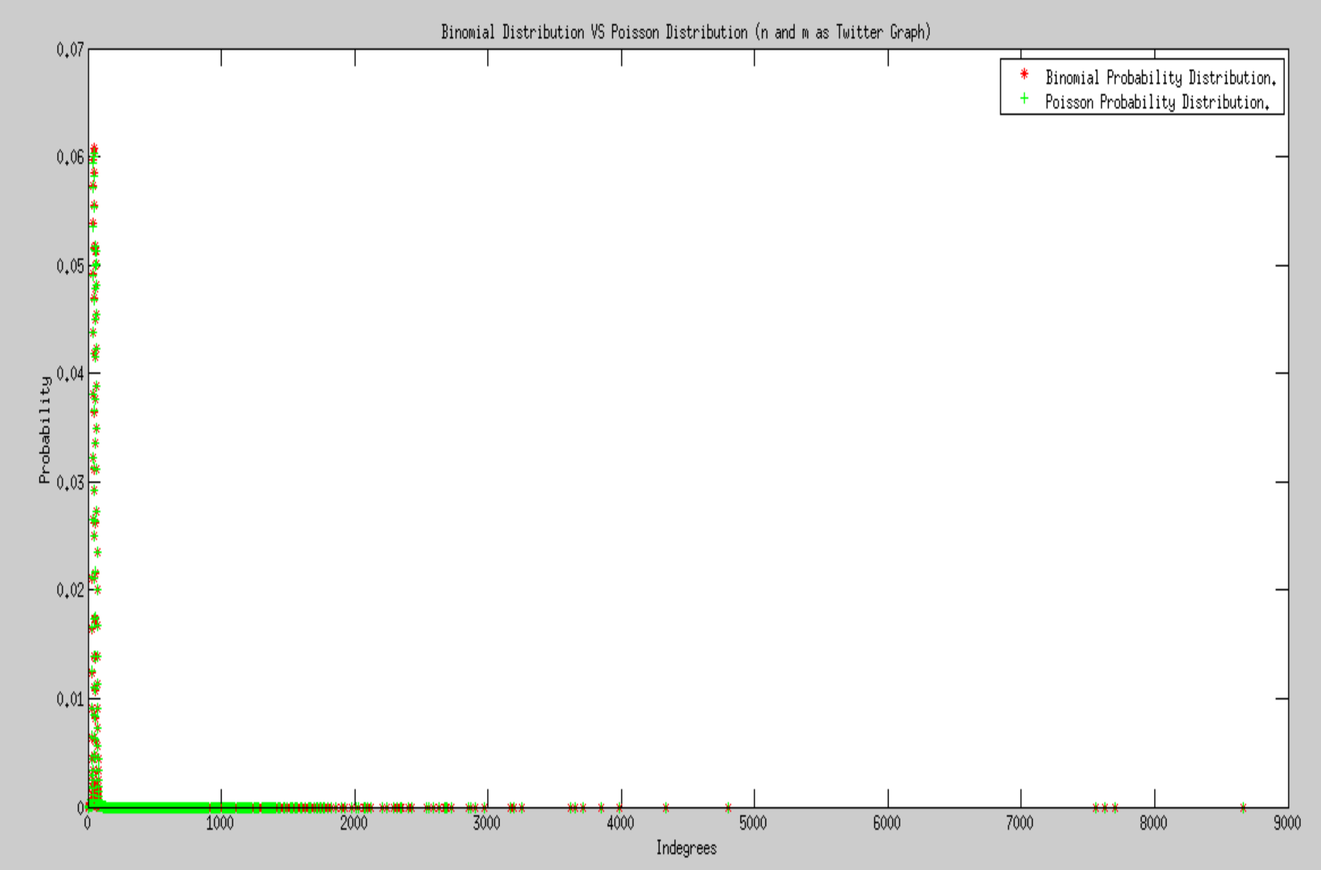
Indegree Calculation

```
select destination_id,count(destination_id) as indegree from twitter_5037 group by(destination_id) order by(destination_id);
```

This query output, which is the indegree for all the nodes in the random graph must be stored in a file so that plotting can be done for the degrees and the binomial and poisson distributions. The following command is used to store the output of the query in a file:

Storing results of query in a file:

```
\COPY (select destination_id,count(destination_id) as indegree from twitter_5037 group by(destination_id) order by(destination_id)) TO '/rmt/csfiles/pgrads/mbva620/Ass3-2_2_1.csv'WITH CSV;
```



The above graph shows the plot between the indegrees in x-axis and the probability (P(k)), which can be Binomial Probability Distribution (or) Poisson Probability Distribution in the y-axis. The red (*) denotes the Binomial Distribution and the green (+) denotes the Poisson Distribution. The comparison between the distribution is done as follows:

Comparison of Results

- The graph clearly shows that the binomial and the poisson distribution looks similar and overlaps at many instances. This is because binomial and poisson distribution are the same but poisson distribution varies with large values of nodes(n). Since the values of n are the same in both the distributions, both binomial and poisson distributions mostly overlap with each other but there are a few differences in the distribution.

- We can observe that binomial distribution has the probability '0' when its degree is 0 and for a few degrees between 1000 and 3000, between 3000 and 4000, at 4400 and 4900, at 7600, 7700, 7800 and finally at the degree 8800 but poisson distribution has the probability 0 at many degrees such as from 0 to 2000 and the values of degrees where the binomial distribution also had probability 0.

- Then almost all the values of binomial distribution overlaps with poisson distribution except a few point differences between the two distributions. For instance, at degree 100, the value of poisson distribution is around 0.052 and it is 0.051 for binomial distribution, which is the highest probability of both binomial and poisson distribution. Similarly, there is only a small difference between the binomial and poisson distribution for each degree and poisson seems to exceed binomial distribution but by a very small value. This is because, as the value of n(total number of nodes) becomes larger, the binomial distribution is approximated as poisson distribution. As n increases, p value decreases and the computation value of ‘lambda’, which is used for the calculation of P(k) in poisson distribution decreases further. In the above graph, the p value was 7.3239e-04 , which gave the high peaks in poisson distribution and the split in binomial distribution.

Outdegree:

Query:

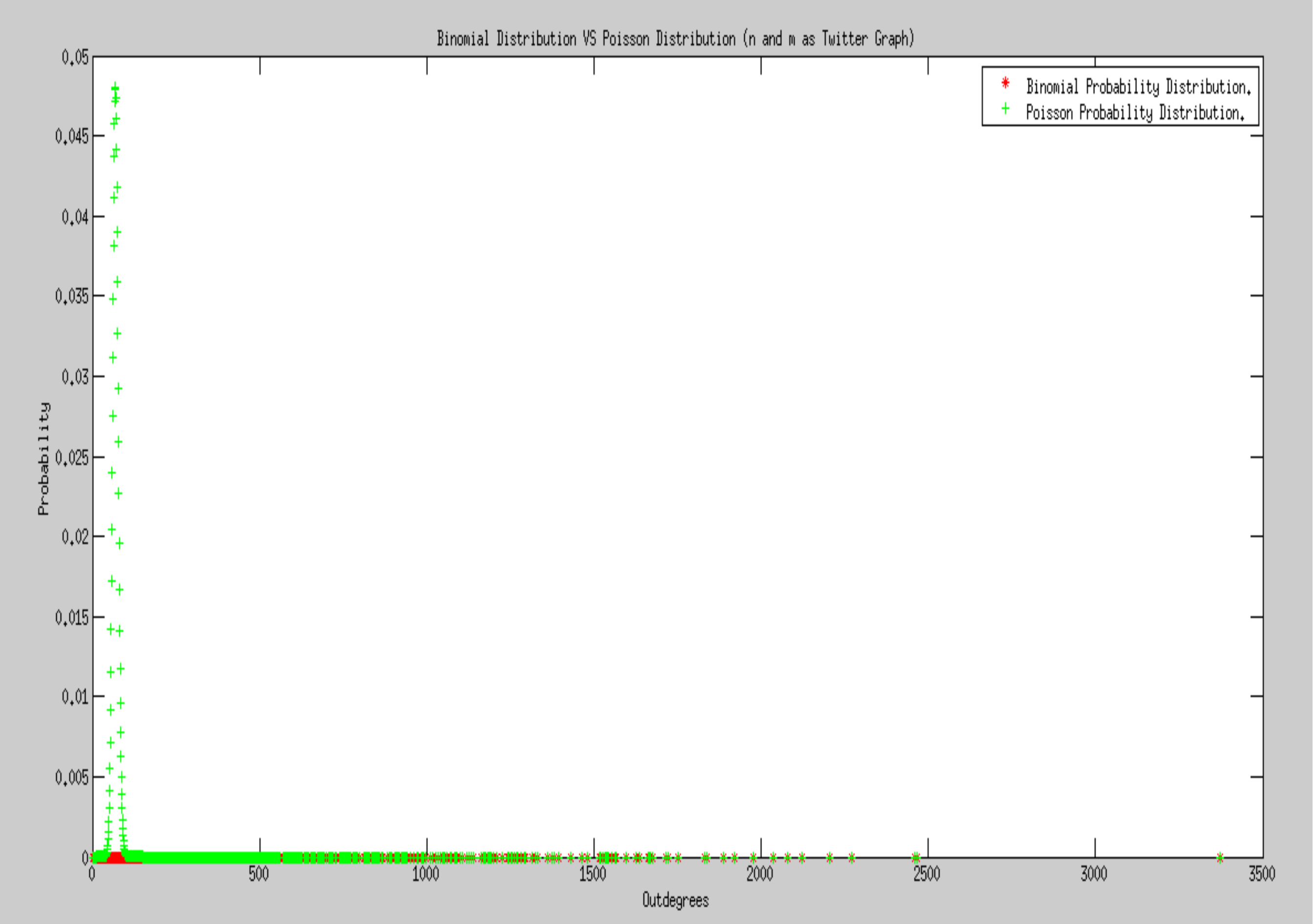
Outdegree Calculation

select source_id,count(source_id) as outdegree from twitter_5037 group by(source_id) order by(source_id);

This query output, which is the outdegree for all the nodes in the random graph must be stored in a file so that plotting can be done for the degrees and the binomial and poisson distributions. The following command is used to store the output of the query in a file:

Storing results of query in a file:

\COPY (select source_id,count(source_id) as outdegree from twitter_5037 group by(source_id) order by(source_id)) TO '/rmt/csfiles/pgrads/mbva620/Ass3-2_2_1_1.csv'WITH CSV;



The above graph shows the plot between the indegrees in x-axis and the probability (P(k)), which can be Binomial Probability Distribution (or) Poisson Probability Distribution in the y-axis. The red (*) denotes the Binomial Distribution and the green (+) denotes the Poisson Distribution. The comparison between the distribution is done as follows:

Comparison of Results

- The graph clearly shows that the binomial and the poisson distribution looks similar and overlaps at many instances from above 500 to 2500 and near 3400. We can observe that both binomial distribution and poisson distribution has the probability '0' when its outdegree is 0 and for a few outdegrees between 1000 and 1500, from 1500 to 2000, from 2000 to 2500, and finally at the degree around 3400 and probability just above 0 at the degrees between 500 to 1000 and for few values from 1000 to around 1250.

- The binomial distribution in the above graph is very low because the n or the total number of nodes is large. So, when the n is large, the binomial distribution is usually approximated using the poisson distribution. Hence, we can observe in the above graph that the peak can be observed in poisson distribution clearly than in binomial distribution. The poisson distribution reaches its maximum value of around 0.048 within few degree and then declines. This is the poisson distribution that can be observed when the p value is less than 0.05, and the p value is 9.8535e-04 and the binomial almost declines for higher values of n.

(2.1.2) Degree Distribution of Actual Twitter Graph

Indegree

The degree distribution of actual twitter graph can be found by taking the indegrees that are present for each node in the twitter graph. Then, taking into account all the indegrees in total, the number of occurrences of each degree is taken as the count. This count is divided by the total number of nodes to find the proportion of the degree occurrence in the twitter graph. This proportion of each degree can be calculated as follows:

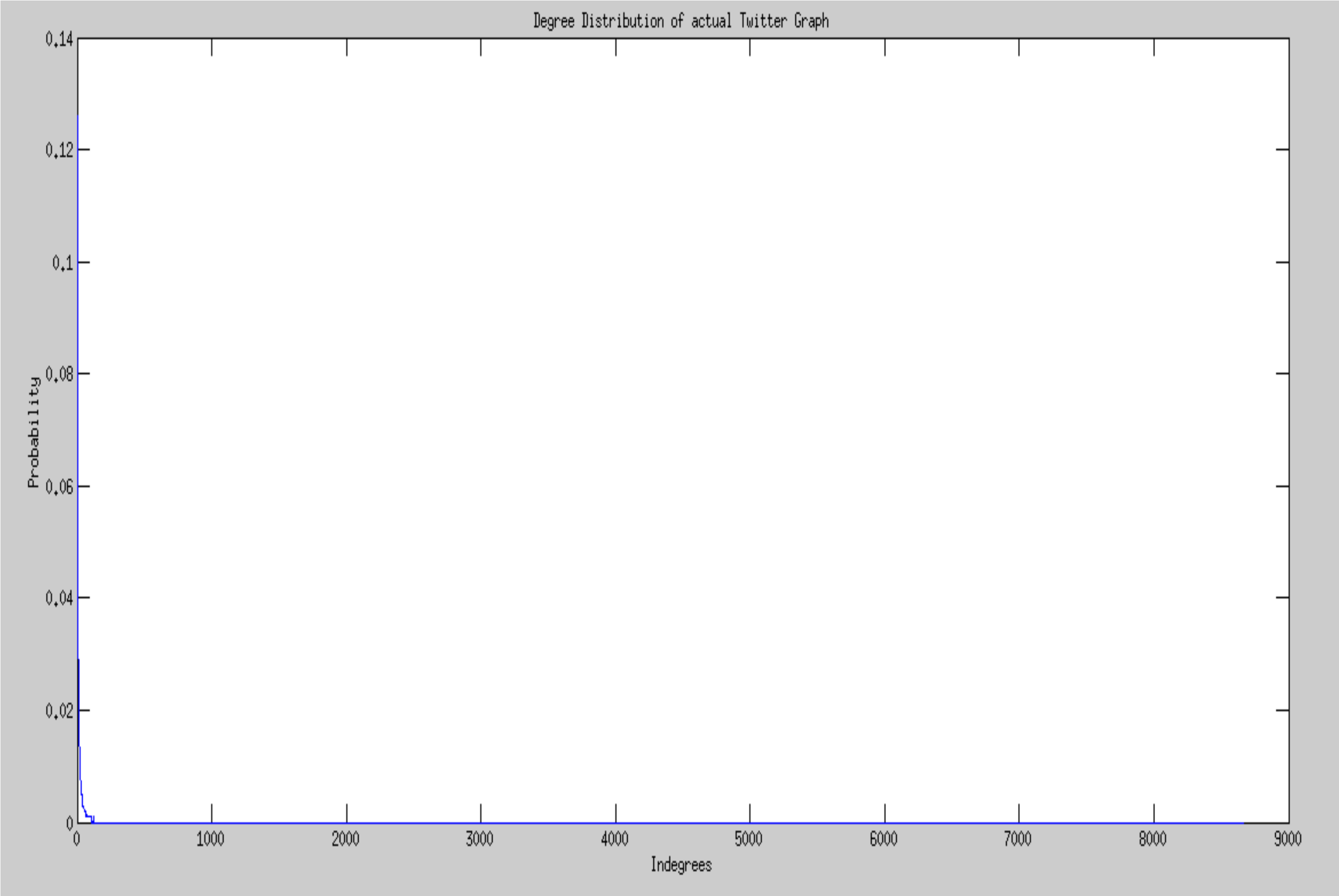
Query:

```
with a as(select destination_id,count(destination_id) as indegree from twitter_5037 group by(destination_id))select indegree,round((count(destination_id)/(with b as(select source_id from twitter_5037 union select destination_id from twitter_5037)select count(distinct source_id) as n from b)::numeric),3)from a group by(indegree) order by(indegree);
```

This query outputs the indegree and the proportion of that indegree occurrence in the twitter graph. This output has to be stored in a file to plot the degree distribution of twitter graph. The following query is used to store the output in a file:

Storing results of query in a file:

```
\COPY (with a as(select destination_id,count(destination_id) as indegree from twitter_5037 group by(destination_id))select indegree,round((count(destination_id)/(with b as(select source_id from twitter_5037 union select destination_id from twitter_5037)select count(distinct source_id) as n from b)::numeric),3)from a group by(indegree) order by(indegree)) TO '/rmt/csfiles/pgrads/mbva620/Ass3-indegree.csv'WITH CSV;
```



The degree distribution of actual twitter graph is formed by plotting the indegrees in the x-axis and the proportion of the indegrees' occurrence, that is, (number of times the indegree has occurred across all the nodes in the graph)/ (total number of nodes(n)).

(2.1.4) Interpretations

Indegree

- The indegrees are distributed in the actual twitter graph is similar to the power law graph, that is, P(k) ~ x^{-alpha}, where alpha < 1, with a long tail. The power law, according to statistics, is a functional relationship between two quantities, where one quantity varies as a power of another. So, the degree distribution of twitter graph looks similar to the linear scale of the Power law distribution. If we plot the degree distribution of twitter graph in log-log scale, we get a straight line which joins both the axes.

- The degree distribution of twitter graph does not exactly match the poisson distribution, whereas it looks similar to a graph plotted by power law distribution. The plot of power law shows that the initial values y-axis values are very high with small values of the term in x-axis and then the value in y-axis decreases gradually with an increase in x-axis values and we can observe that the above graph looks similar to the power law distribution. Hence, the plot on degree distribution of twitter graph shows that the shape of this degree distribution matches most with the power law distribution.

- Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The above graph shows a positive skew or right skewed, that is, there is a long tail on the right side of the graph. The right tail is longer; the mass of the distribution is concentrated on the left of the figure. So, larger share of population rests within its tail, that is, many degrees are covered (from around 50 to around 8700) by the long tail on the right side of the graph. We can also observe that these degrees on the long tail have the probability 0 in the degree distribution and also there are many nodes with the degree 0, as mostly the proportion of nodes with the degree 0. Since the degree distribution is positively skewed, the mean is greater than the median reflecting the fact that the mean is sensitive to each degree in the distribution and is subject to large shifts when the sample is small and contains extreme degrees.

- There are a few hubs in the twitter graph, that is, the nodes which are exceptionally well connected to the other nodes in the graph. The above degree distribution of twitter graph is a plot between the indegrees and the proportion of nodes of each degree (count of each indegree / total number of nodes). So, it can be observed from the above graph that the many nodes had the degree around 0 and then it declined from 1 to about 50. The node which has the highest indegree is the hub as it will be well connected among various nodes in the graph, such a node is ‘40981798’ which has the highest indegree of 8660. The other hubs or well-connected nodes are 22462180, 34428380, 43003845 with the indegrees as 7623, 7558 and 7700 respectively.

Outdegree

The degree distribution of actual twitter graph can be found by taking the outdegrees that are present for each node in the twitter graph. Then, taking into account all the outdegrees in total, the number of occurrences of each degree is taken as the count. This count is divided by the total number of nodes to find the proportion of the degree occurrence in the twitter graph. This proportion of each degree can be calculated as follows:

Query:

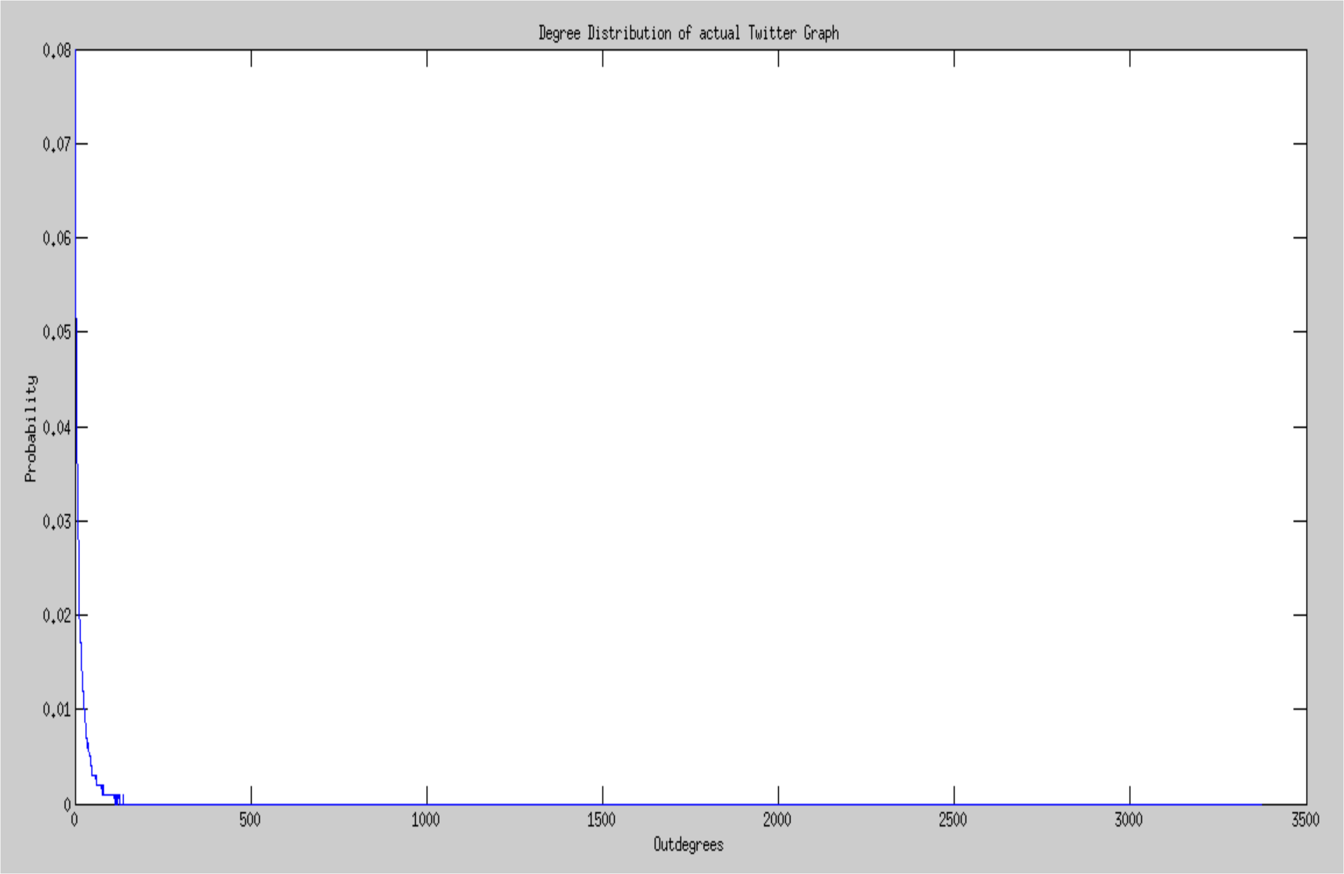
```
with a as(select source_id,count(source_id) as outdegree from twitter_5037 group by(source_id))select outdegree,round((count(source_id)/(with b as(select source_id from twitter_5037 union select destination_id from twitter_5037)select count(distinct source_id) as n from b)::numeric),3)from a group by(outdegree) order by(outdegree);
```

This query outputs the outdegree and the proportion of that outdegree occurrence in the twitter graph. This output has to be stored in a file to plot the degree distribution of twitter graph. The following query is used to store the output in a file:

Storing results of query in a file:

```
\COPY (with a as(select source_id,count(source_id) as outdegree from twitter_5037 group by(source_id))select outdegree,round((count(source_id)/(with b as(select source_id from twitter_5037 union select destination_id from twitter_5037)select count(distinct source_id)
```

as n from b)::numeric),3)from a group by(outdegree) order by(outdegree)) TO '/rmt/csfiles/pgrads/mbva620/Ass3-outdegree.csv'WITH CSV;



The degree distribution of actual twitter graph is formed by plotting the outdegrees in the x-axis and the proportion of the outdegrees' occurrence, that is, (number of times the outdegree has occurred across all the nodes in the graph)/ (total number of nodes(n)).

(2.1.4) Interpretations

Outdegree

- The degrees are distributed in the actual twitter graph is similar to the power law graph, that is, P(k) ~ x^{-alpha}, where alpha < 1, with a long tail. The power law, according to statistics, is a functional relationship between two quantities, where one quantity varies as a power of another. So, the degree distribution of twitter graph looks similar to the linear scale of the Power law distribution and so if we plot the degree distribution of twitter graph in log-log scale, then it is almost a diagonal line connecting the x and the y-axis.

- The degree distribution of facebook graph does not exactly match the poisson distribution, whereas it looks similar to a graph plotted by power law distribution. When power law is plotted, the initial probability values y-axis values are very high with small values of the term in x-axis and then the value in y-axis decreases gradually with an increase in x-axis values. Hence, the above graph on degree distribution of facebook graph shows that the shape of this degree distribution matches most with the power law distribution.

- Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The above graph shows a positive skew or right skewed, that is, there is a long tail on the right side of the graph. The right tail is longer; the mass of the distribution is concentrated on the left of the figure. So, larger share of population rests within its tail, that is, many degrees are covered (from around 200 to around 3400) by the long tail on the right side of the graph. We can also observe that these degrees on the long tail have the probability 0 in the degree distribution. We can also observe that these degrees on the long tail have the probability 0 in the degree distribution and also there are many nodes with the degree 0, as mostly the proportion of nodes exists with degree 0. Then there were a few nodes above the degree 0 and then this declined when the degree reached around 200. Since the degree distribution is positively skewed, the mean is greater than the median reflecting the fact that the mean is sensitive to each degree in the distribution and is subject to large shifts when the sample is small and contains extreme degrees.

- There are a few hubs in the twitter graph, that is, the nodes which are exceptionally well connected to the other nodes in the graph. The above degree distribution of twitter graph is a plot between the outdegrees and the proportion of nodes of each outdegree (count of each outdegree / total number of nodes). So, it can be observed from the above graph that the many nodes had the outdegree around 0 to 200 and then it declined. The node which has the highest outdegree is the hub as it will be well connected among various nodes in the graph, such a node is '3359851' which has the highest degree of 3373. The other hubs or well- connected nodes are 5442012, 7860742, 16098603,18776017,59804598,83943787 and 270449528 which the next few higher outdegrees.

(3) <u>Questions</u>

<u>3.1 Properties of Social Network</u>

<u>New Properties of Social Network:</u>

- **<u>Clustering Interest:</u>** When a user have an interest at a particular topic, the probability that the user will have interest in similar topics.
- **<u>Assigning Priority to friends:</u>** Some friends can be assigned as best friends or can be considered as favourites, so they are given the first priority and even groups can be formed in social and information can be shared only among those groups, this shows that filtration can also be done among friends.
- **<u>Hiding data for security:</u>** Some confidential information can be focussed for a particular set of users of the social network. So, hiding of information from all the users, that is filtering the viewers of the information.

- **Spatial Properties:** There lies heterogeneity in the characteristic distance of interaction across users. Some preference is given to short-range rather than long-distance ties.

We have discussed about the following properties of social network in this assignment:

Transitivity: Social ties between A and B and between B and C might imply a tie between A and C

Clustering coefficient: Probability of an edge between A and C, given edges between A and B and between B and C, averaged over all nodes

Betweenness: How many of the shortest paths between people in the graph pass through a particular person X.

Funneling: Most information passes through a few people

3.2 Social Network becomes matured

When social network is first formed, it does not have many nodes as well many edges but as it grows, many people come to know that they can interact through social network and the number of nodes, that is, the number of people and edges, that is, the number of connection between people increase. So, when the social network matures, the number of nodes and edges also increases rapidly. For example, the social networks like twitter, facebook didn’t have more than 4000 users or 4000 nodes during their initial stages, but now there are billions of users of twitter and facebook and as the nodes increases, the connection between the nodes also increases which leads to the increase in the number of edges. Hence, the data storage for social network also increases, small data storage units are not sufficient to manage the data. For example, our facebook dataset has about 4000 users or nodes which is very less when compared to the original facebook dataset. This can be managed by leveraging big data tools as the data will be growing enormously in the social network as it matures.

3.3 Ego-network

Friendship graph is a graph formed for a particular Facebook user , where the user’s friends are only taken into consideration and the linkage between them are only considered.

This can be calculated like what we did to find the actual number of triangles in the Facebook graph:

```
select count(*) as actual_num_of_traingles from facebook t1,facebook t2,facebook t3 where
t1.destination_id=t2.source_id and t2.destination_id=t3.destination_id and
t1.source_id=t3.source_id;
```

This query finds the number of friends who form a triangle. Let us consider our Facebook user as the primary source node or source_id , that is, t1.source_id and is compared with his connections in the destination_id's and whether they are connected to each other. For example, A is connected to B, and A is connected to C, we have to check whether B is connected to C and the above query gives that count.

The friendship graph drawn for a user who does not travel frequently and from a small town

If the Facebook user does not travel frequently and is also from a small town, then his connections will be less. So, the destination_id that the primary source_id has will be less and also we have to check whether those destination_id are connected to each other, the probability of their connection will be very less. So, in this case, the connections will be less and the graph won't be populated with many nodes being connected through edges.

The friendship graph drawn for a user who travels frequently and knew many groups of people

If the Facebook user travel frequently and knew many groups of people, the probability that those groups of people being friends is very high. Travel makes people get more connections, since our Facebook user travels frequently, he might have new connections and since it also says that our Facebook user knew many groups of people, the probability that people belonging to one group might know each other is very high and also some people who belong to different groups might be friends with each other. So, in this case, the connections will be very high and the graph will be populated with many nodes and many edges connecting these nodes.

3.4 Identifying rogue Facebook account

The rogue facebook account can be found by various methods:

- If that rogue facebook user gives a friend request and you come to know that the same facebook user has given request to all your friends who belong to different groups, then you can identify that the rogue Facebook user tries to befriend everyone.
- If the rogue Facebook user has connections with your friends who basically belong to different groups and are not connected in any way to each other, then you can confirm that the rogue Facebook user is trying to be friends with all Facebook users
- You can view the timeline of the rogue Facebook user, if none of the information about the user is clear or seems suspicious, then we can confirm that person as a rogue Facebook user.

(4) Six Degrees of Separation

The n value indicates the number of hops taken by the nodes. Within 3 hops, we try to check, whether a particular node is linked with all the other nodes in the facebook graph. This analysis is made on each node that is present in the facebook fraph, that is, whether each node is connected to every other node in a maximum of 3 hops. We use the following query to find those degree nodes for a particular source node:

n=3

```
with t1 as(select source_id,destination_id from facebook_5037 union select
destination_id,source_id from facebook_5037), t2 as(select a.source_id as
src_id,b.destination_id as snd_deg from t1 a,t1 b where a.destination_id = b.source_id
and b.destination_id <> a.source_id except select * from t1),s2 as(select src_id,
snd_deg from t2 where snd_deg in (select snd_deg from t2 intersect select source_id
from t1)), s3 as(select source_id,destination_id from t1 where source_id in (select
snd_deg from t2 intersect select source_id from t1))select s2.src_id,s3.destination_id
from s2 join s3 on s2.snd_deg=s3.source_id except select * from t1 except select *
from t2;
```

n=2

```
with t1 as(select source_id,destination_id from facebook_5037 union select
destination_id,source_id from facebook_5037)select a.source_id as
src_id,b.destination_id as snd_deg from t1 a,t1 b where a.destination_id = b.source_id
and b.destination_id <> a.source_id except select * from t1;
```

n=1

```
select source_id,destination_id from facebook_5037 union select
destination_id,source_id from facebook_5037;
```

But, the output of these queries does not lead to the connection of each node to all the other nodes in the graph. So, we have to consider higher degrees such as n=4, n=5 or n=6 to check whether each node of the graph is connected to the all the other nodes.

(5) Conclusion

The social network provides a medium for various users or nodes and these various users are connected through edges. We can observe from the comparison and interpretation above that the data in social network is growing and the data storage units are increasing as the number of users is being increased, the possibility of mutual connections and new connections also increases. The various distributions show the proportion of users of social network is becoming high and hence the normal probability distributions are not enough to analyse the data. We can use various big data tools to analyse the data of the social networks.