Social Network Analysis Assignment

Time-Aware Network Centrality Measures & Link Prediction

Argyropoulou Meropi BAFT1722

Tsopelakou Aglaia BAFT 1716

Lisa Stephanie Schramm BAFT1701

Contents

[**PART I** 2](#_Toc523704112)

[**PART II** 9](#_Toc523704113)

Table of Figures

[**Figure 1: Subgraph G0** 3](#_Toc523705613)

[**Figure 2:Subgraph G1** 3](#_Toc523705614)

[**Figure 3: Time Evolution of Vertices and Edges** 4](#_Toc523705615)

[**Figure 4: Probability Density Function-Degree Centrality-SubGraph G0** 5](#_Toc523705616)

[**Figure 5: Probability Density Function-Degree Centrality-SubGraph G69** 6](#_Toc523705617)

[**Figure 6: Probability Density Function-Closeness Centrality-SubGraph G0** 6](#_Toc523705618)

[**Figure 7: Probability Density Function-Betweeness Centrality- SubGraph G0** 7](#_Toc523705619)

[**Figure 8: Probability Density Function-Eigen Vector Centrality- SubGraph G0** 7](#_Toc523705620)

[**Figure 9: Probability Density Function-Eigen Vector Centrality- SubGraph G69** 8](#_Toc523705621)

[**Figure 10: Probability Density Function-Katz Centrality- SubGraph G0** 8](#_Toc523705622)

[**Figure 11: Time Evolution of V, E\*prev, E\*after** 12](#_Toc523705623)

[**Figure 12: Similarity values Common Neighbours** 14](#_Toc523705624)

Table of Tables

[**Table 1: Time Periods** 2](#_Toc523705607)

[**Table 2: |V \*[tj-1, tj+1]|** 9](#_Toc523705608)

[**Table 3: |E\*[tj-1, tj ]|** 10](#_Toc523705609)

[**Table 4: |E\*[tj , tj+1]|** 11](#_Toc523705610)

[**Table 5: Similarity Measures example** 13](#_Toc523705611)

# **PART I**

1. Partition the complete time period T = [tmin, tmax] into a set of non-overlapping time periods {T1, : : : , TN} by computing the corresponding set of time instances {t0, …, tN} where t­0= tmin and tN = tmax. Mind that N is a user defined parameter.

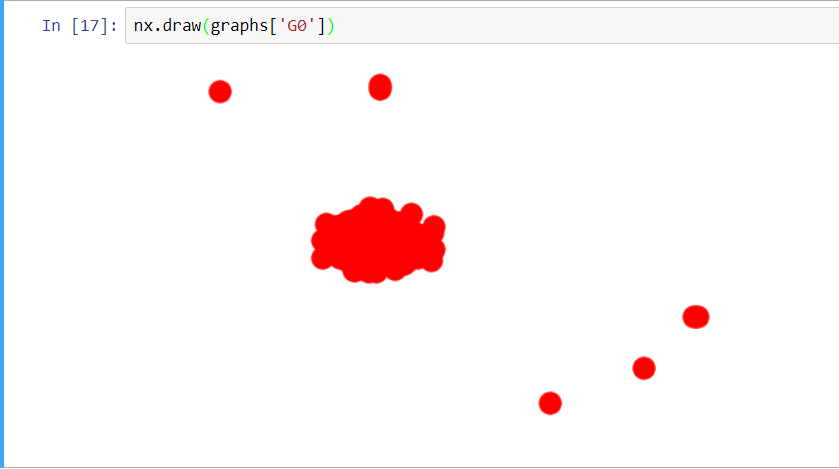
We partitioned the dataset into N=91 time periods shown in the Table 1 starting from 31/08/2008 @ 5:00pm (UTC) and ending in 06/03/2016 @ 2:10pm (UTC). Later in our analysis due to memory restrictions of our PCs (8 GB RAM) we created the sub-networks for the 70 time periods out of the 91, until the 05/06/2014 8:26. Although we tried we different values for N, the network increases dramatically (1 million nodes for -N=91-) and cumulatively (more subnetworks with a large number of nodes -for N=2500-).

**Table 1: Time Periods**

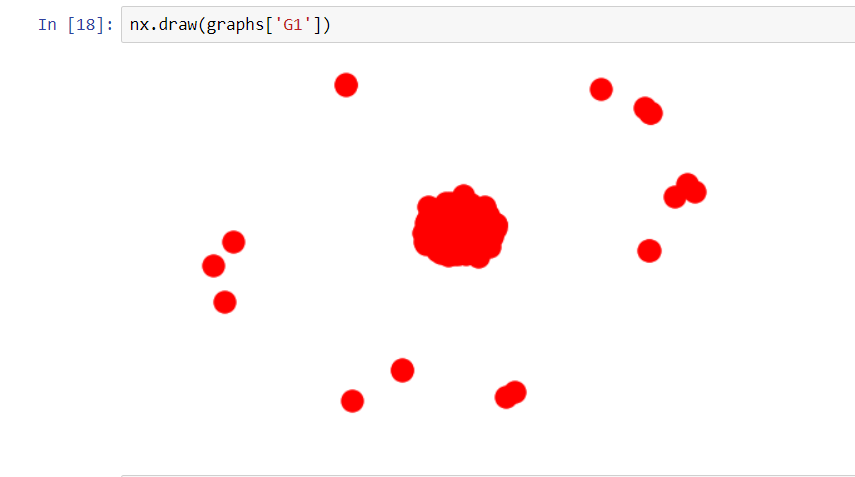
|  |  |  |
| --- | --- | --- |
| 31/08/2008 17:00 | 03/04/2011 19:45 | 03/11/2013 22:31 |
| 01/10/2008 4:42 | 04/05/2011 7:27 | 04/12/2013 10:13 |
| 31/10/2008 16:24 | 03/06/2011 19:09 | 03/01/2014 21:55 |
| 01/12/2008 4:06 | 04/07/2011 6:52 | 03/02/2014 9:37 |
| 31/12/2008 15:48 | 03/08/2011 18:34 | 05/03/2014 21:19 |
| 31/01/2009 3:30 | 03/09/2011 6:16 | 05/04/2014 9:01 |
| 02/03/2009 15:12 | 03/10/2011 17:58 | 05/05/2014 20:43 |
| 02/04/2009 2:54 | 03/11/2011 5:40 | 05/06/2014 8:26 |
| 02/05/2009 14:36 | 03/12/2011 17:22 | 05/07/2014 20:08 |
| 02/06/2009 2:19 | 03/01/2012 5:04 | 05/08/2014 7:50 |
| 02/07/2009 14:01 | 02/02/2012 16:46 | 04/09/2014 19:32 |
| 02/08/2009 1:43 | 04/03/2012 4:28 | 05/10/2014 7:14 |
| 01/09/2009 13:25 | 03/04/2012 16:11 | 04/11/2014 18:56 |
| 02/10/2009 1:07 | 04/05/2012 3:53 | 05/12/2014 6:38 |
| 01/11/2009 12:49 | 03/06/2012 15:35 | 04/01/2015 18:20 |
| 02/12/2009 0:31 | 04/07/2012 3:17 | 04/02/2015 6:02 |
| 01/01/2010 12:13 | 03/08/2012 14:59 | 06/03/2015 17:45 |
| 31/01/2010 23:56 | 03/09/2012 2:41 | 06/04/2015 5:27 |
| 03/03/2010 11:38 | 03/10/2012 14:23 | 06/05/2015 17:09 |
| 02/04/2010 23:20 | 03/11/2012 2:05 | 06/06/2015 4:51 |
| 03/05/2010 11:02 | 03/12/2012 13:47 | 06/07/2015 16:33 |
| 02/06/2010 22:44 | 03/01/2013 1:30 | 06/08/2015 4:15 |
| 03/07/2010 10:26 | 02/02/2013 13:12 | 05/09/2015 15:57 |
| 02/08/2010 22:08 | 05/03/2013 0:54 | 06/10/2015 3:39 |
| 02/09/2010 9:50 | 04/04/2013 12:36 | 05/11/2015 15:22 |
| 02/10/2010 21:32 | 05/05/2013 0:18 | 06/12/2015 3:04 |
| 02/11/2010 9:15 | 04/06/2013 12:00 | 05/01/2016 14:46 |
| 02/12/2010 20:57 | 04/07/2013 23:42 | 05/02/2016 2:28 |
| 02/01/2011 8:39 | 04/08/2013 11:24 | 06/03/2016 14:10 |
| 01/02/2011 20:21 | 03/09/2013 23:07 |  |
| 04/03/2011 8:03 | 04/10/2013 10:49 |

1. Choose an appropriate representation for each subgraph G[t­­­­­­j-1, t­­j ] of the network for each time period Tj where 1 < j < N.

In order to represent the subgraph, we used the python library NetworkX. Below, we can see the representation of the subgraph G0 and the representation of the successive subgraph G1.



**Figure 1: Subgraph G0**

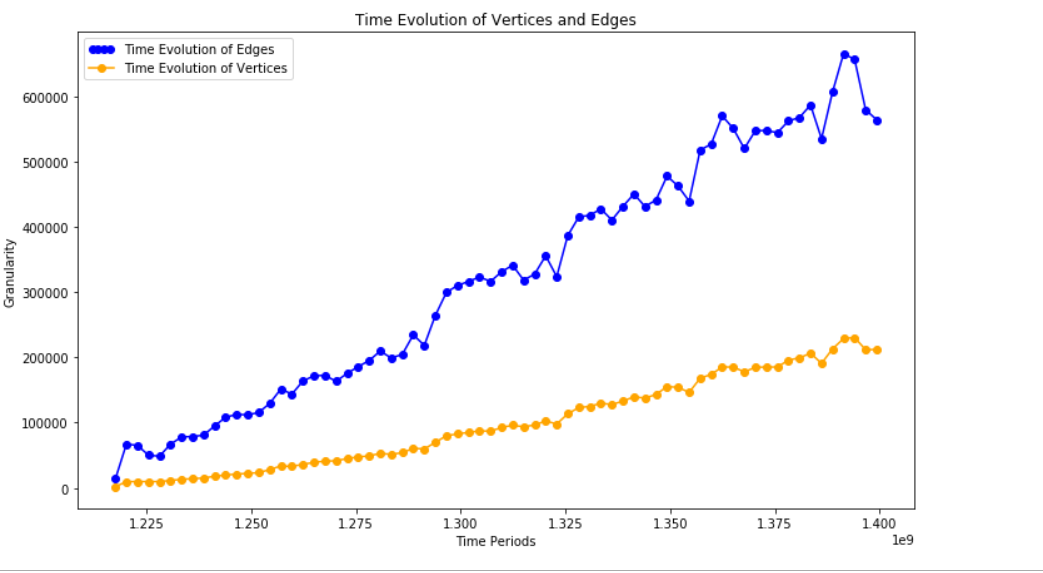


**Figure 2:Subgraph G1**

1. Provide a graph depicting the time evolution of the quantities IV [tj-1, tj ]I and IE[tj-1,tj ]I for each time period Tj where 1 <j< N.

Below in Figure 3 the time evolution of vertices and edges is depicted. The granularity of vertices is slowly rising. The trend of the granularity of edges follows the same pattern but over time is steeper and has a lot of ups and downs especially at the last periods under consideration.

It is expected to have more edges than vertices as the interactions among the nodes is growing larger as the time passes. We might think for the explanation of the ups and downs to correlate them with the fact that when a new technology is released there is a peak in the interactions of the network (questions and answers) but after sometime this drops slightly till a new technology is released.



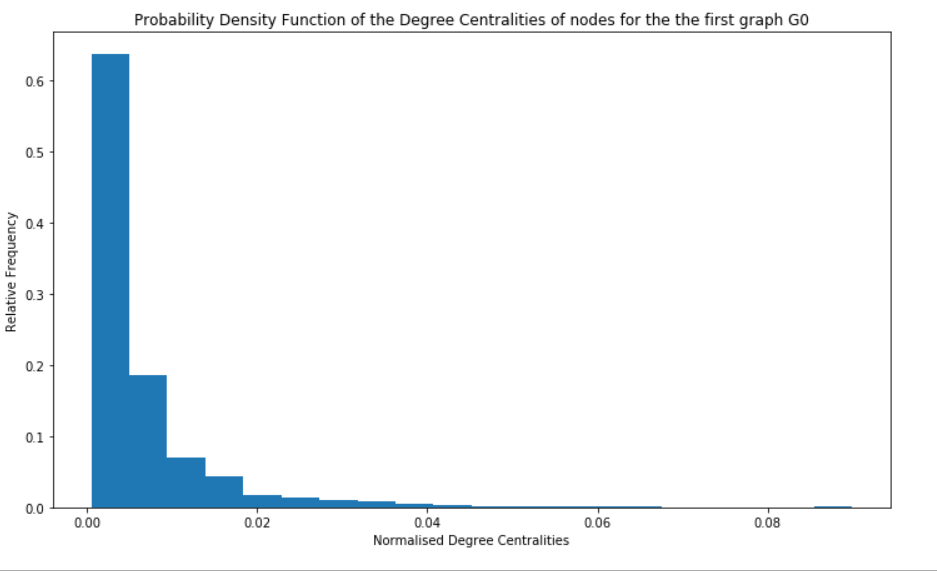
**Figure 3: Time Evolution of Vertices and Edges**

1. For each subgraph G[tj-1, tj ] compute and graphically represent the probability density functions (i.e. histograms of relative frequencies) for the following centrality measures:

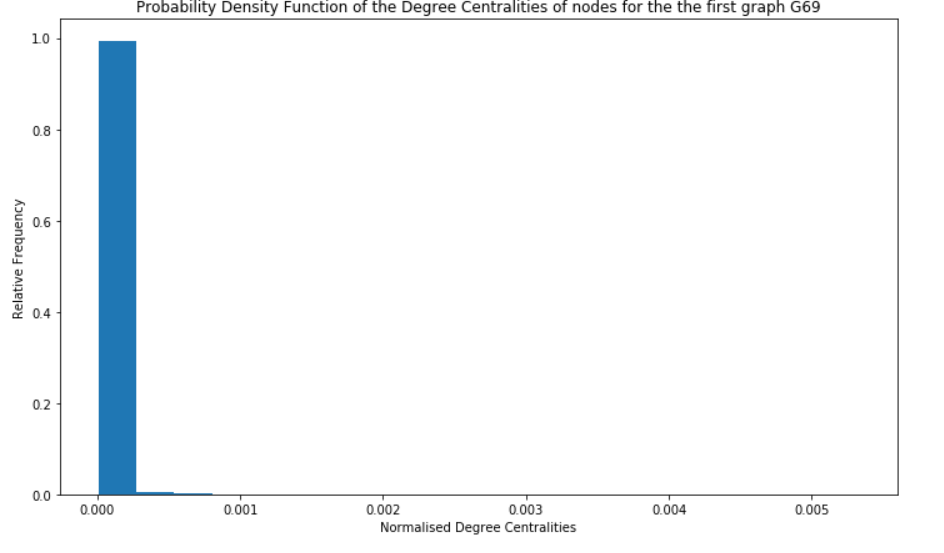
Below are shown histograms of the relative frequencies of the subgraphs for the centralities. For all the centralities there are the histograms of relative frequencies for the first subgraph G0 and for some centralities where we did not get a memory error there are the histograms of relative frequencies of G69 which is the last subgraph.

1. Degree Centrality

For the degree centrality we can see that for G0 subgraph the majority of the nodes had normalized degree centrality close to 0 and some of them they had normalized degree between 0.01 and 0.07. On the contrary, for the subgraph G69 almost all the nodes had normalized degree centrality close to 0. That means that the nodes in G0 graph had more connections(comments) than those of G69 probably because as the time passes there are more comments that do not get answer in comparison with the beginning of Stack Overflow.



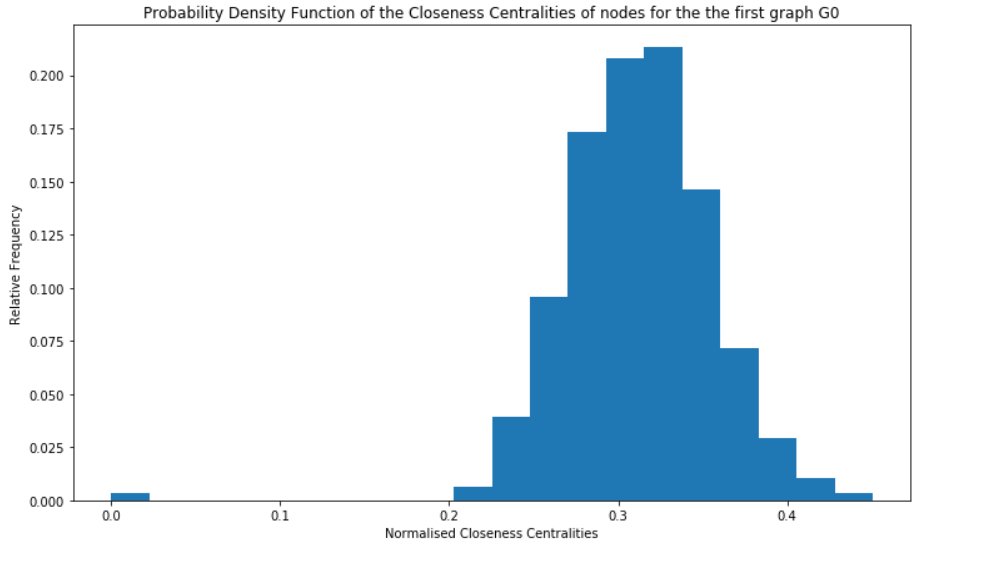
**Figure 4: Probability Density Function-Degree Centrality-SubGraph G0**



**Figure 5: Probability Density Function-Degree Centrality-SubGraph G69**

1. Closeness Centrality

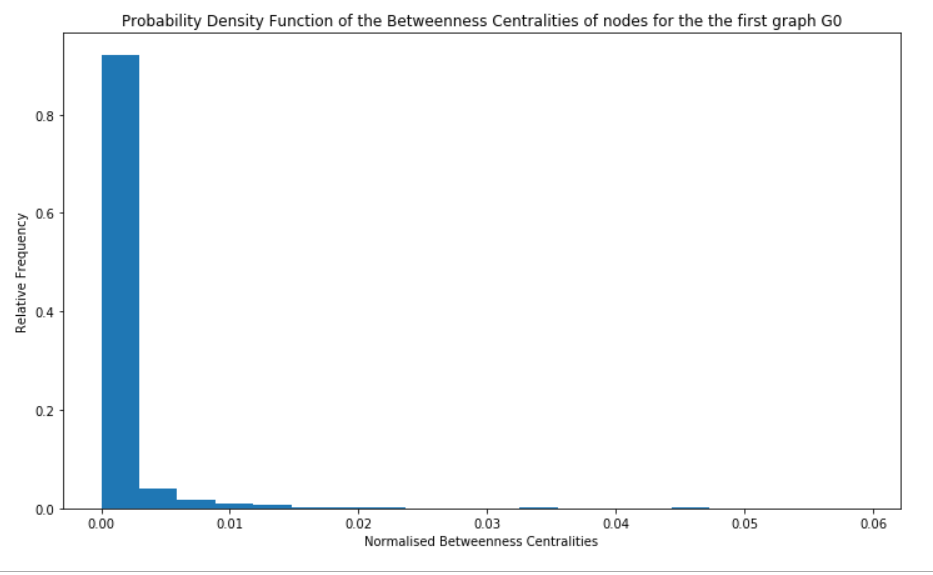
Below the frequency of G0 is represented. The normalized closeness centrality of G0 subgraph follows almost a normal distribution(bell curve distribution) starting from 0.2 to 0.5 with higher rank in 0.3. Subgraph G69 could not be loaded due to memory restrictions.



**Figure 6: Probability Density Function-Closeness Centrality-SubGraph G0**

1. Betweenness Centrality

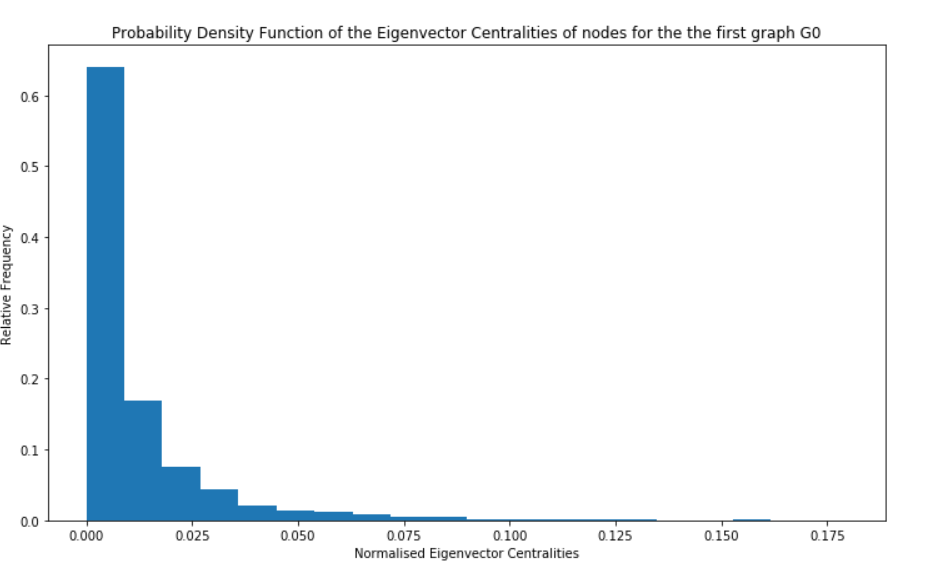
The subgraph G0 is listed below. As we can see we have really low normalized betweenness centrality almost 0 for all of our nodes. That means that we do not have any nodes with considerable influence within the network by virtue of their control over information passing between others.



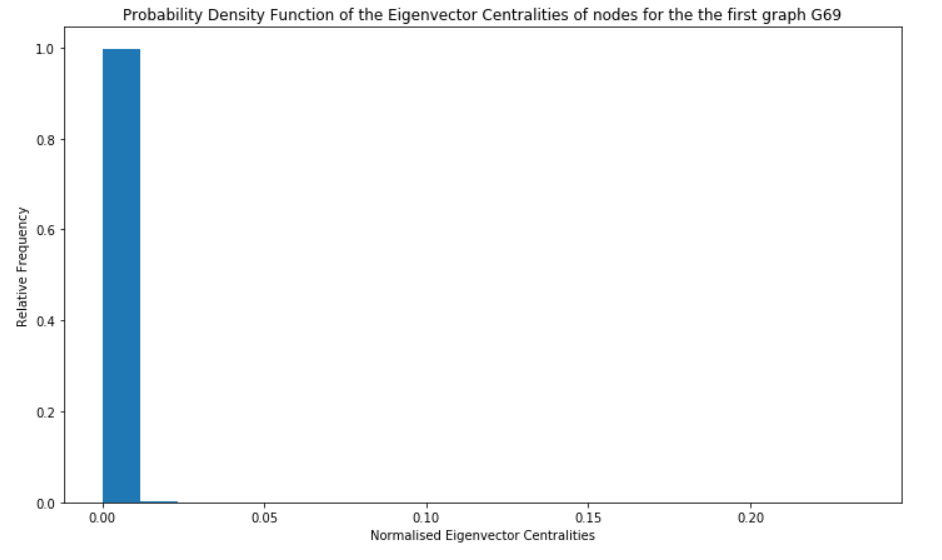
**Figure 7: Probability Density Function-Betweeness Centrality- SubGraph G0**

1. Eigenvector Centrality

For eigenvector centrality for G0 subgraph the distribution is power-law on the contrary with G69 graph where almost all the nodes have normalized eigenvector centrality 0. However, we notice that for both of the subgraphs, really low values eigenvector centralities exist. Thus, nodes are not connected to many nodes who themselves have high scores.



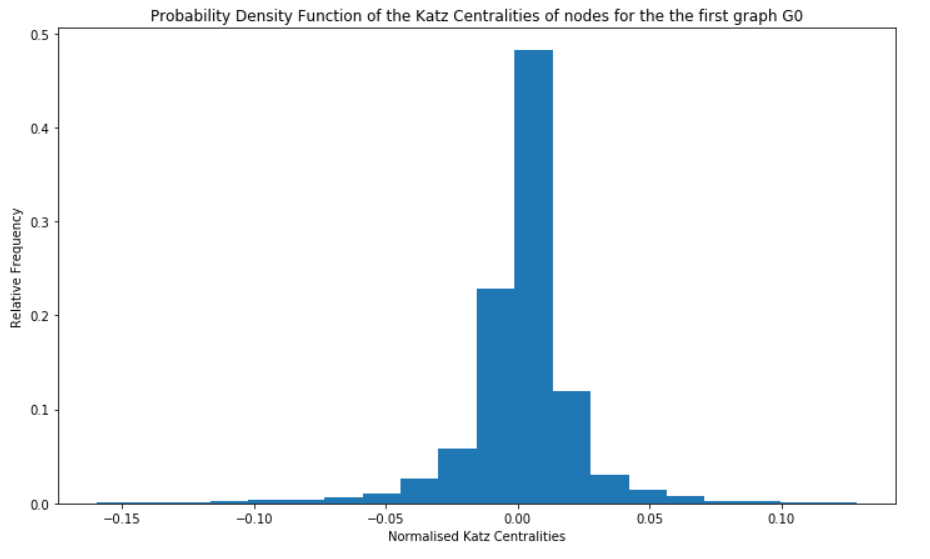
**Figure 8: Probability Density Function-Eigen Vector Centrality- SubGraph G0**



**Figure 9: Probability Density Function-Eigen Vector Centrality- SubGraph G69**

1. Katz Centrality

Because our subgraphs are not (strongly) connected we can use Katz to solve the previous problem with eigenvector centrality. In the figure below, we can see that the Katz centrality for subgraph G0 follows a bell distribution starting from -0.05 and ending in 0.10.



**Figure 10: Probability Density Function-Katz Centrality- SubGraph G0**

# **PART II**

1. For each pair of successive network instances (G[tj-1, tj ],G[tj , tj+1]), where1 < j <N - 1, compute the following sets:

(a) V \*[tj-1, tj+1]

(b) E\*[tj-1, tj ]

(c) E\*[tj , tj+1]

and graphically represent their volumes IV \*[tj-1, tj+1]I, IE\*[tj-1, tj ]I and IE\*[tj , tj+1]I as functions of the coupled time periods (Tj , Tj+1).

We are interested in capturing network evolution of the common nodes V\* of two successive time periods, let’s say T1 and T2. An example of common nodes is when a user has made a comment to another user in both sequential time periods T1 and T2 so they have common nodes. Below, all the granularities of the common nodes are listed from period 1-2 (1735 common nodes) through period 69-70 (100292 common nodes).

**Table 2: |V \*[tj-1, tj+1]|**

|  |  |  |
| --- | --- | --- |
| 1735 | 22631 | 60119 |
| 6006 | 23934 | 59969 |
| 5970 | 24341 | 62622 |
| 5829 | 24663 | 63380 |
| 6323 | 26528 | 63416 |
| 7460 | 27769 | 67808 |
| 8308 | 29396 | 69484 |
| 8696 | 34577 | 66918 |
| 9687 | 37879 | 69450 |
| 11208 | 38450 | 76423 |
| 12105 | 39884 | 80068 |
| 12938 | 40277 | 82643 |
| 13242 | 41156 | 79077 |
| 14226 | 43159 | 78925 |
| 15912 | 43243 | 80306 |
| 16681 | 43771 | 80729 |
| 17083 | 46948 | 82188 |
| 18850 | 46729 | 85877 |
| 19900 | 48822 | 87789 |
| 19991 | 55758 | 85097 |
| 20496 | 58557 | 85908 |
| 21669 | 60154 | 96225 |
|  |  | 100292 |

Then, the E\*[tj-1, tj ] are these edges of the first time period let’s say T1 whose both nodes belong to the intersection of let’s say T1-T2. Below the respective volumes (volumes\_of\_edges\_before\_star) through periods T1-T70.

**Table 3: |E\*[tj-1, tj ]|**

|  |  |
| --- | --- |
| 12239 | 203380 |
| 50394 | 201889 |
| 48654 | 193891 |
| 36269 | 207413 |
| 37761 | 214462 |
| 52941 | 204677 |
| 60217 | 243876 |
| 58948 | 254559 |
| 62002 | 256168 |
| 72265 | 252067 |
| 81603 | 242230 |
| 84355 | 256657 |
| 82008 | 259623 |
| 87142 | 248425 |
| 94698 | 259658 |
| 101776 | 267558 |
| 97633 | 249986 |
| 115280 | 250756 |
| 117438 | 292615 |
| 113250 | 299127 |
| 109530 | 312574 |
| 115057 | 288554 |
| 121511 | 278831 |
| 128065 | 292108 |
| 132072 | 291795 |
| 129146 | 290413 |
| 134776 | 299033 |
| 147785 | 300329 |
| 142246 | 292110 |
| 175421 | 279361 |
| 190871 | 324715 |
| 195339 | 351116 |
| 197347 | 325894 |
| 199733 | 281856 |
| 197334 |  |

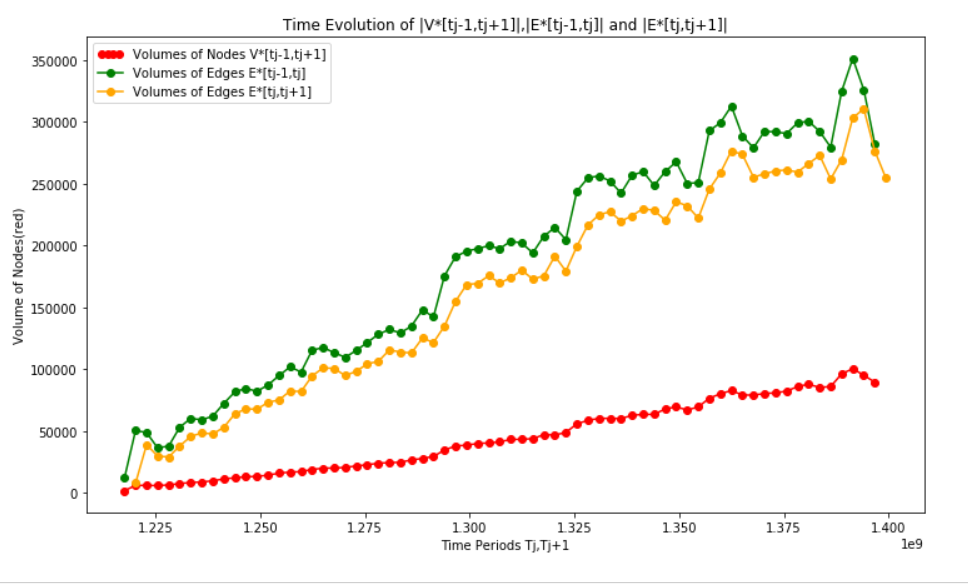
Afterwards, E\*[tj , tj+1] are these edges of the second successive time period let’s say T2 whose both nodes belong to the intersection of let’s say T1-T2. Below the respective volumes (volumes\_of\_edges\_after\_star) through periods T1-T70.

**Table 4: |E\*[tj , tj+1]|**

|  |  |
| --- | --- |
| 7874 | 179945 |
| 38855 | 172736 |
| 29647 | 175342 |
| 28938 | 191326 |
| 37858 | 179164 |
| 45509 | 199272 |
| 48506 | 216512 |
| 47470 | 224699 |
| 52990 | 227258 |
| 63815 | 219605 |
| 67903 | 223891 |
| 67546 | 229891 |
| 73102 | 228136 |
| 75058 | 220393 |
| 82249 | 235323 |
| 81866 | 231910 |
| 94455 | 222066 |
| 101007 | 245462 |
| 100524 | 258504 |
| 94713 | 276425 |
| 98223 | 273503 |
| 104216 | 255205 |
| 106233 | 257781 |
| 115431 | 260202 |
| 113803 | 260947 |
| 112973 | 259230 |
| 125091 | 266091 |
| 121453 | 272765 |
| 134824 | 253679 |
| 154770 | 269335 |
| 168300 | 303031 |
| 169246 | 310424 |
| 175471 | 276304 |
| 169675 | 254454 |
| 173722 |  |

So, when volumes of edges after star is 7874 means that there are 7874 edges from the T2 time period where their nodes belong to the intersection with the nodes of the T1 time period.

Finally, we constructed a diagram with the volume of nodes, the evolution of the volume of edges of T1 and the volume of edges of T2. In this diagram, we can see that the volume of nodes, the volume of edges T1 and the volume of edges T2 are increasing over time. Of course, the nodes are less that the edges but the edges of the E\*T1 are more than the edges of the E\*T2 even though they are proportional. Also, as the time goes by at the end we can see some ups and downs in both of the edges as expected.



**Figure 11: Time Evolution of V, E\*prev, E\*after**

1. For each pair of nodes (u, v) ∉ V \*[tj-1, tj+1] and for every set of common vertices V \*[tj-1, tj+1], where 1 <j <N - 1, compute the following similarity matrices:

(a) SGD : [Graph Distance]

(b) SCN : [Common Neighbors]

(c) SJC : [Jaccard's Coefficient]

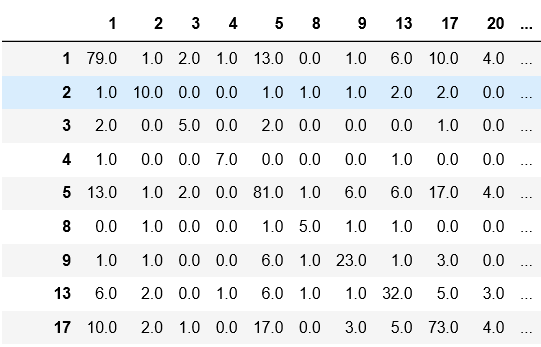
(d) SA : [Adamic / Adar]

(e) SPA : [Preferential Attachment]

*\*\*Please scroll down and ignore the <class 'networkx.exception.NetworkXError'> printed in the code. It was to handle the case where the pair was not existent in the network. It was printed to monitor whether the exception was handled. After this there is also code for the rest of the measures.*

For each common set of vertices (intersection) between two successive time periods, we created a dataframe V\* **\*** V\* with indices these values and filledin with 0s. Then for all the possible pairs we calculated the similarity values.

In the code you can see in the respective part the matrix (dataframe) of all the **Common Neighbours** similarity values for all the possible pairs of the vertices V\*[t1,t2] in the E\* graph created from the edges of the V\*[t1,t2] that exist in the intersection.



**Figure 12: Similarity values Common Neighbours**

Similar matrices can be created for the rest of the measures. Below an example of the values of the five similarity measures for the pair (1,2).

**Table 5: Similarity Measures example**

|  |  |
| --- | --- |
| **Similarity Measure** | **Result for pair (1,2)** |
| Graph Distance | 1 |
| Common Neighbors | 1 |
| Jaccard's Coefficient | 0.011 |
| Adamic / Adar | 0.22 |
| Preferential Attachment | 810 |