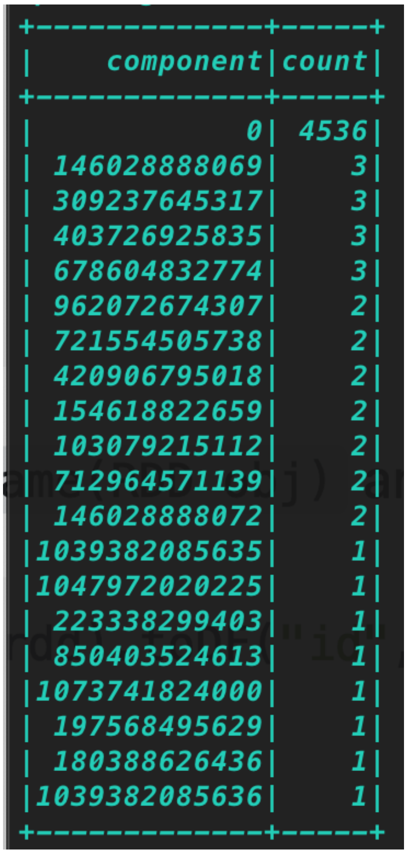
IS459 BDA Assignment 2

Question 1

How large are the communities (connected components)?

The graph used here to analyze the connected components is formed by connecting the nodes of each author (source author) with the node of another author (destination author) that they interacted with on the HardwareZone PC Gaming Forum. As a form of data pre-processing, I've also removed the edges where the source author = destination author, which was created when an author comments on his/her own post *(lines 54-58 of assignment_2.py)*.

By applying the connectedComponents() function on the graph, we notice that out of 4661 distinct authors on HardwareZone, 4536 of them are interconnected with each other, and there are many other smaller connected components with 3 or less authors in it, as seen in this DataFrame below. *(lines 75-80 of assignment_2.py)*



```
+-------------+-----+
|    component|count|
+-------------+-----+
|            0| 4536|
| 146028888069|    3|
| 309237645317|    3|
| 403726925835|    3|
| 678604832774|    3|
| 962072674307|    2|
| 721554505738|    2|
| 420906795018|    2|
| 154618822659|    2|
| 103079215112|    2|
| 712964571139|    2|
| 146028888072|    2|
|1039382085635|    1|
|1047972020225|    1|
| 223338299403|    1|
| 850403524613|    1|
|1073741824000|    1|
| 197568495629|    1|
| 180388626436|    1|
|1039382085636|    1|
+-------------+-----+
```

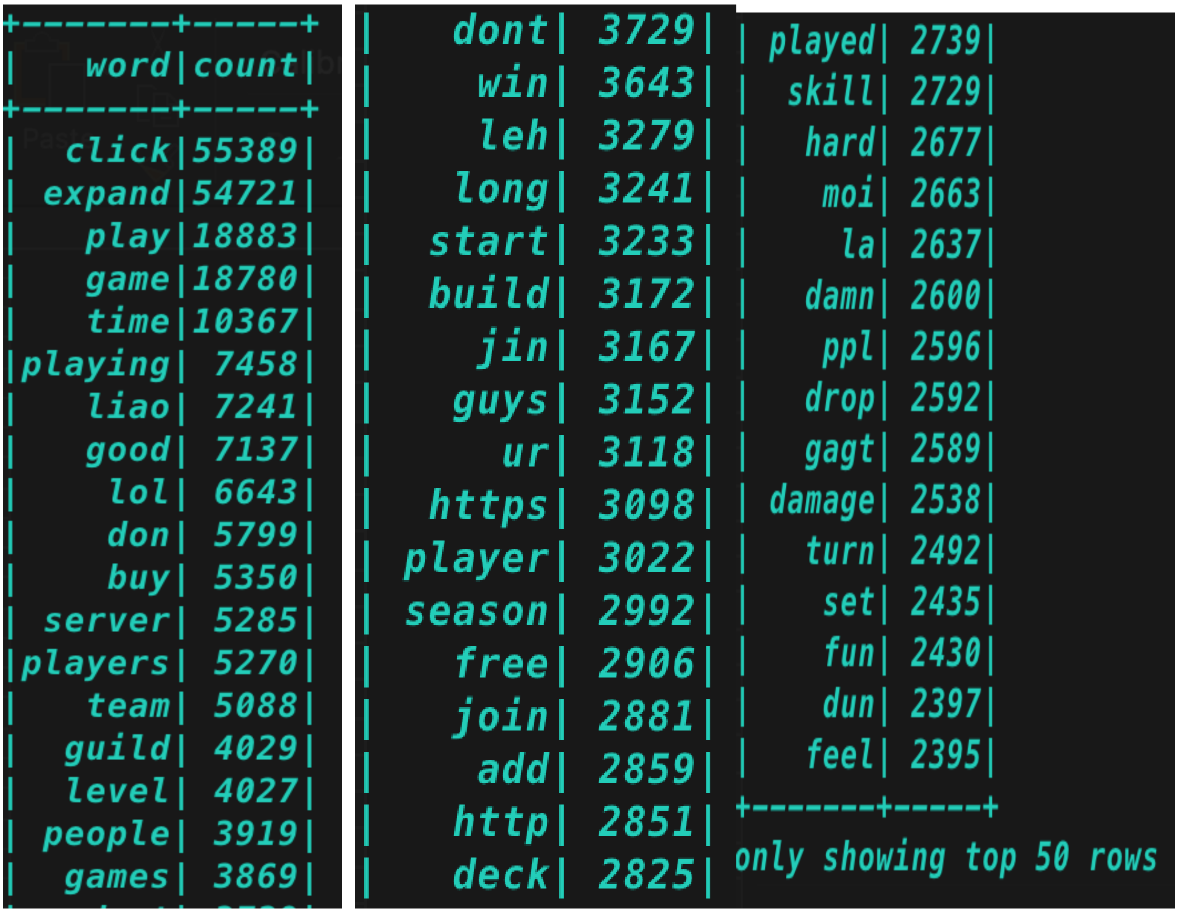
As such, I would conclude that in general, the 4661 authors on HardwareZone are rather **closely connected with one another** as a majority (4536 authors) are connected to one another in one single connected component. There are a total of **111** connected components in the community, where **110** of these connected components contain **less than or equal to 3 authors**.

What are the key words of the community (frequent words)?

In order to get the frequent keywords, I’ve retrieved a list of stopwords from https://countwordsfree.com/stopwords and downloaded it as a text file, and uploaded it to HDFS so that I’m able to read it on assignment_2.py. By tokenizing every word in content and removing it if it’s a stopword, I’ve gathered the following list of common keywords, sorted in descending order.

Attached below is a list of the top 50 keywords, sorted in descending order.

(lines 82-96 of assignment_2.py)

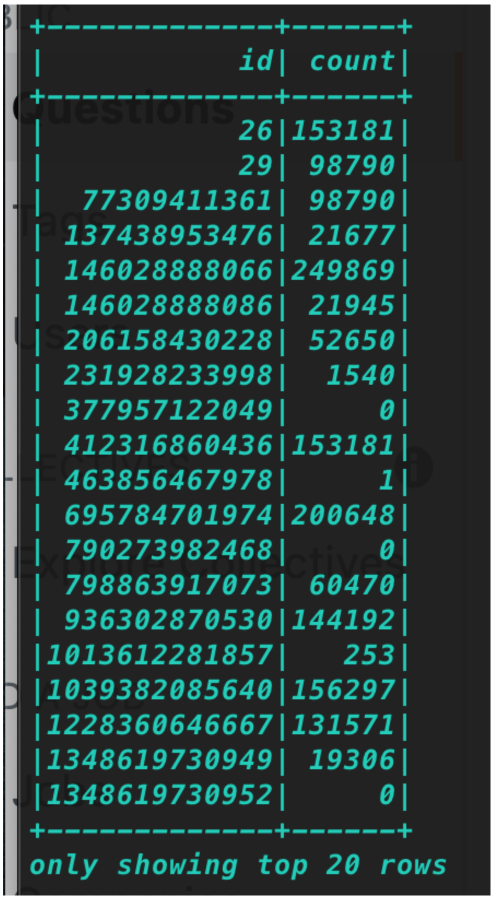


```
+-------+-----+
|   word|count|
+-------+-----+
|  click|55389|
| expand|54721|
|   play|18883|
|   game|18780|
|   time|10367|
|playing| 7458|
|   liao| 7241|
|   good| 7137|
|    lol| 6643|
|    don| 5799|
|    buy| 5350|
| server| 5285|
|players| 5270|
|   team| 5088|
|  guild| 4029|
|  level| 4027|
| people| 3919|
|  games| 3869|
|   dont| 3729|
|    win| 3643|
|    leh| 3279|
|   long| 3241|
|  start| 3233|
|  build| 3172|
|    jin| 3167|
|   guys| 3152|
|     ur| 3118|
|  https| 3098|
| player| 3022|
| season| 2992|
|   free| 2906|
|   join| 2881|
|    add| 2859|
|   http| 2851|
|   deck| 2825|
| played| 2739|
|  skill| 2729|
|   hard| 2677|
|    moi| 2663|
|     la| 2637|
|   damn| 2600|
|    ppl| 2596|
|   drop| 2592|
|   gagt| 2589|
| damage| 2538|
|   turn| 2492|
|    set| 2435|
|    fun| 2430|
|    dun| 2397|
|   feel| 2395|
+-------+-----+
only showing top 50 rows
```

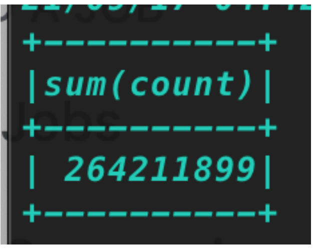
Question 2

How cohesive are the communities (Average # of triangles over every user in a community)?

(lines 101-106 of assignment_2.py) By applying the triangleCount() function on the graph, we're able to get the total number of triangles formed for each author. We can then sum up the number of triangles, which is **264211899**.

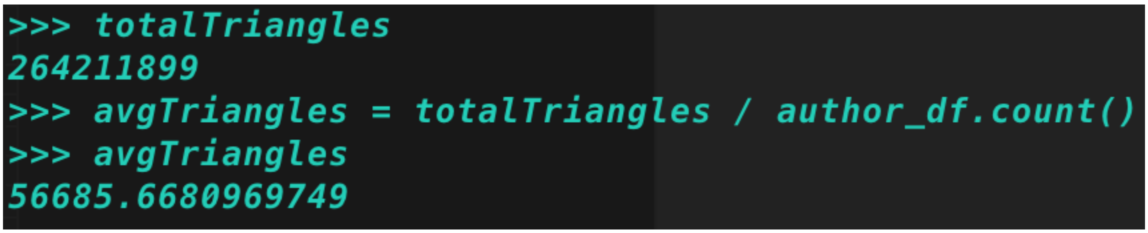


```
+-------------+------+
|           id| count|
+-------------+------+
|           26|153181|
|           29| 98790|
|  77309411361| 98790|
| 137438953476| 21677|
| 146028888066|249869|
| 146028888086| 21945|
| 206158430228| 52650|
| 231928233998|  1540|
| 377957122049|     0|
| 412316860436|153181|
| 463856467978|     1|
| 695784701974|200648|
| 790273982468|     0|
| 798863917073| 60470|
| 936302870530|144192|
|1013612281857|   253|
|1039382085640|156297|
|1228360646667|131571|
|1348619730949| 19306|
|1348619730952|     0|
+-------------+------+
only showing top 20 rows
```



```
+----------+
|sum(count)|
+----------+
| 264211899|
+----------+
```

In order to determine the cohesiveness of the community, we can then find out the average number of triangles across all the authors on the HardwareZone PC Gaming Forum, which is an average of **56685** triangles. Hence, we can conclude that the community in the HardwareZone PC Gaming Forum is indeed cohesive.



```
>>> totalTriangles
264211899
>>> avgTriangles = totalTriangles / author_df.count()
>>> avgTriangles
56685.6680969749
```

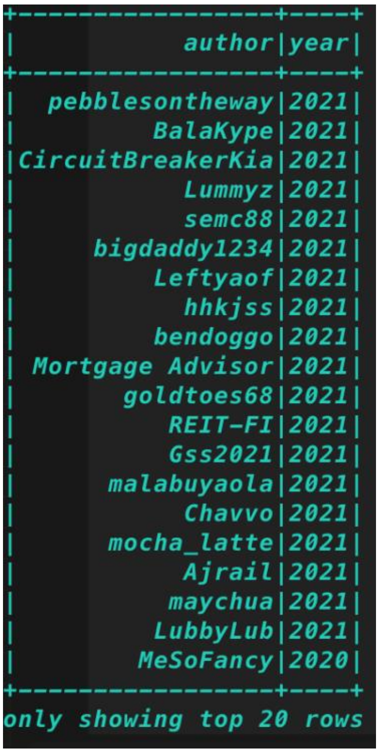
Question 3

Is there any strange community? (Open) – E.g., younger generation or older generation?

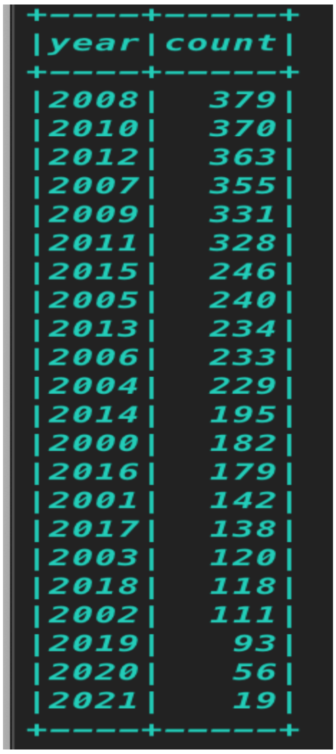
(lines 111-120 of assignment_2.py) For this question, I've crawled additional data of the Join Date of every author on HardwareZone. The Join Date can also be a good gauge of the "age" of the author on the HardwareZone PC Gaming Forum, and for this question, we'll be making the following assumptions:

- An author is considered younger generation if he/she joined after 2011 (Year >= 2011)
- An author is considered older generation if he/she joined before 2011 (Year < 2011)

First, we can gather the data and create a dataframe as seen below (left image), with a data frame of the author and the year they've joined HardwareZone.

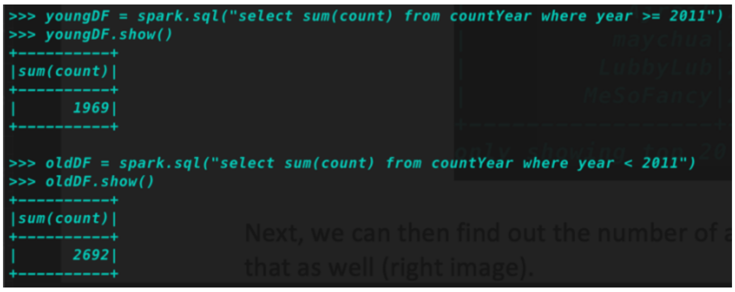


```
+------------------+----+
|            author|year|
+------------------+----+
|   pebblesontheway|2021|
|          BalaKype|2021|
|CircuitBreakerKia|2021|
|            Lummyz|2021|
|            semc88|2021|
|       bigdaddy1234|2021|
|          Leftyaof|2021|
|            hhkjss|2021|
|          bendoggo|2021|
|  Mortgage Advisor|2021|
|        goldtoes68|2021|
|           REIT-FI|2021|
|           Gss2021|2021|
|        malabuyaola|2021|
|            Chavvo|2021|
|        mocha_latte|2021|
|            Ajrail|2021|
|           maychua|2021|
|          LubbyLub|2021|
|         MeSoFancy|2020|
+------------------+----+
only showing top 20 rows
```



```
+----+-----+
|year|count|
+----+-----+
|2008|  379|
|2010|  370|
|2012|  363|
|2007|  355|
|2009|  331|
|2011|  328|
|2015|  246|
|2005|  240|
|2013|  234|
|2006|  233|
|2004|  229|
|2014|  195|
|2000|  182|
|2016|  179|
|2001|  142|
|2017|  138|
|2003|  120|
|2018|  118|
|2002|  111|
|2019|   93|
|2020|   56|
|2021|   19|
+----+-----+
```

Next, we can then find out the number of authors who joined every year and create a dataframe for that as well (right image).



```
>>> youngDF = spark.sql("select sum(count) from countYear where year >= 2011")
>>> youngDF.show()
+----------+
|sum(count)|
+----------+
|      1969|
+----------+

>>> oldDF = spark.sql("select sum(count) from countYear where year < 2011")
>>> oldDF.show()
+----------+
|sum(count)|
+----------+
|      2692|
+----------+
```

From the image above, we can tell that there's actually more authors under the older generation category **(2692 authors**), as compared to the younger generation category (**1969 authors**). Hence, I'll conclude that there's **more authors from the older generation**, based on the assumption mentioned above.