【Experiment name】 Identification and Analysis of Influential Developers in Open Source Software Community

【Purpose】

Understanding the Implications of Influencer Developer Identification

Understand the various indicators and their meanings that characterize the importance of network nodes

Master the basic principles of PageRank algorithm

Master the method of GraphFrames to build a network and related network index calculation interface

Master the network visualization tool Gephi

【Experimental content】

Topic :

Use GraphFrames to analyze the attention relationship network among PHP developers in GitHub, find developers with important influence, and conduct multi-dimensional analysis and visualization.

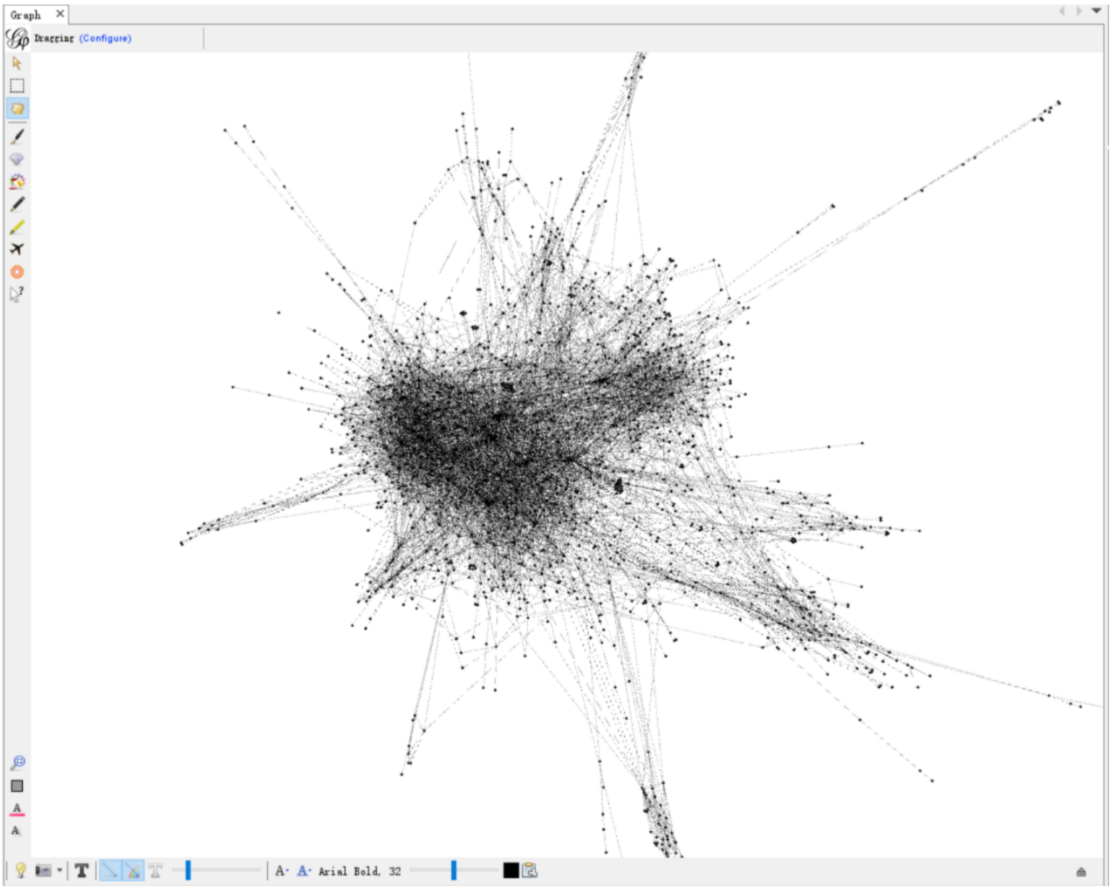
Requirements :

1. Configure the GraphFrames environment;
2. Use Gephi to visualize the developer's attention relationship network;
3. Use GraphFrames to calculate the developer's PageRank value, and obtain the information of the top 10 developers ;
4. Conduct multi-dimensional analysis on the top 10 influential developers, including their out-degree, in-degree, and number of triangles, etc.

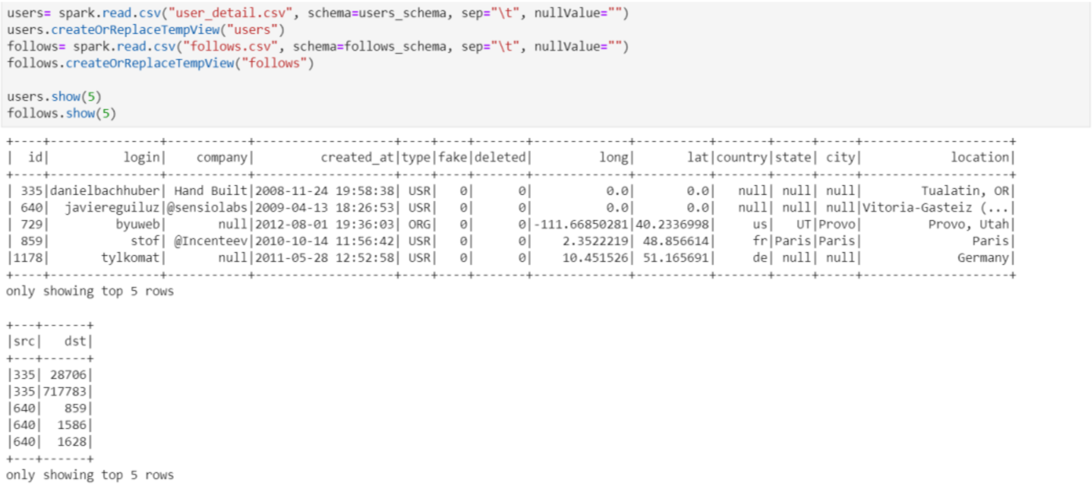
Improve (optional) :

Experimental results (experimental steps and related core codes) :

The relationship network diagram generated by Gephi :



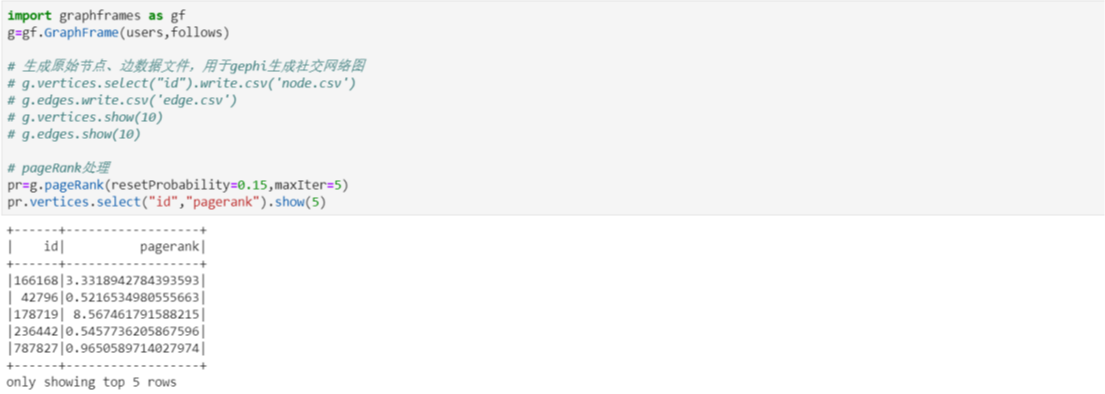
Experimental code analysis:



```
users= spark.read.csv("user_detail.csv", schema=users_schema, sep="\t", nullValue="")
users.createOrReplaceTempView("users")
follows= spark.read.csv("follows.csv", schema=follows_schema, sep="\t", nullValue="")
follows.createOrReplaceTempView("follows")

users.show(5)
follows.show(5)
```

```
+----+---------------+-----------+-------------------+----+----+-------+-------------+----------+-------+-----+-----+--------------------+
|  id|          login|    company|         created_at|type|fake|deleted|         long|       lat|country|state| city|            location|
+----+---------------+-----------+-------------------+----+----+-------+-------------+----------+-------+-----+-----+--------------------+
| 335|danielbachhuber| Hand Built|2008-11-24 19:58:38| USR|   0|      0|          0.0|       0.0|   null| null| null|        Tualatin, OR|
| 640|  javiereguiluz|@sensiolabs|2009-04-13 18:26:53| USR|   0|      0|          0.0|       0.0|   null| null| null|Vitoria-Gasteiz (...|
| 729|         byuweb|       null|2012-08-01 19:36:03| ORG|   0|      0|-111.66850281|40.2336998|     us|   UT|Provo|         Provo, Utah|
| 859|           stof| @Incenteev|2010-10-14 11:56:42| USR|   0|      0|    2.3522219| 48.856614|     fr|Paris|Paris|               Paris|
|1178|       tylkomat|       null|2011-05-28 12:52:58| USR|   0|      0|    10.451526| 51.165691|     de| null| null|             Germany|
+----+---------------+-----------+-------------------+----+----+-------+-------------+----------+-------+-----+-----+--------------------+
only showing top 5 rows

+---+------+
|src|   dst|
+---+------+
|335| 28706|
|335|717783|
|640|   859|
|640|  1586|
|640|  1628|
+---+------+
only showing top 5 rows
```

① Read the data file, there are users and f ollows respectively



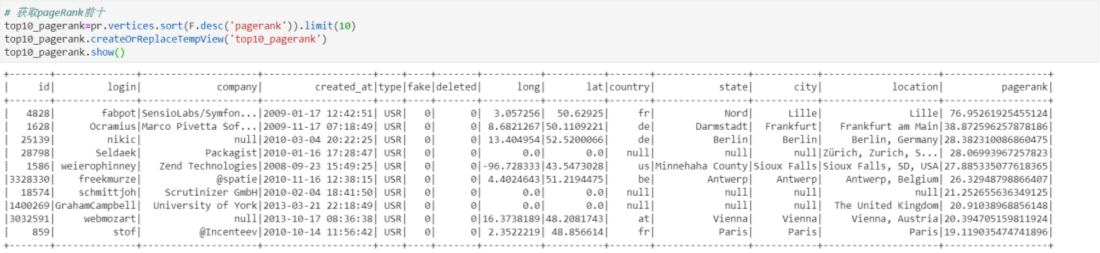
```
import graphframes as gf
g=gf.GraphFrame(users,follows)

# 生成原始节点、边数据文件，用于gephi生成社交网络图
# g.vertices.select("id").write.csv('node.csv')
# g.edges.write.csv('edge.csv')
# g.vertices.show(10)
# g.edges.show(10)

# pageRank处理
pr=g.pageRank(resetProbability=0.15,maxIter=5)
pr.vertices.select("id","pagerank").show(5)
```

```
+------+------------------+
|    id|          pagerank|
+------+------------------+
|166168|3.3318942784393593|
| 42796|0.5216534980555663|
|178719| 8.567461791588215|
|236442|0.5457736205867596|
|787827|0.9650589714027974|
+------+------------------+
only showing top 5 rows
```

② Do PageRank processing according to the requirements of s slide

* 1:(resetProbability=0.15,maxIter=5)

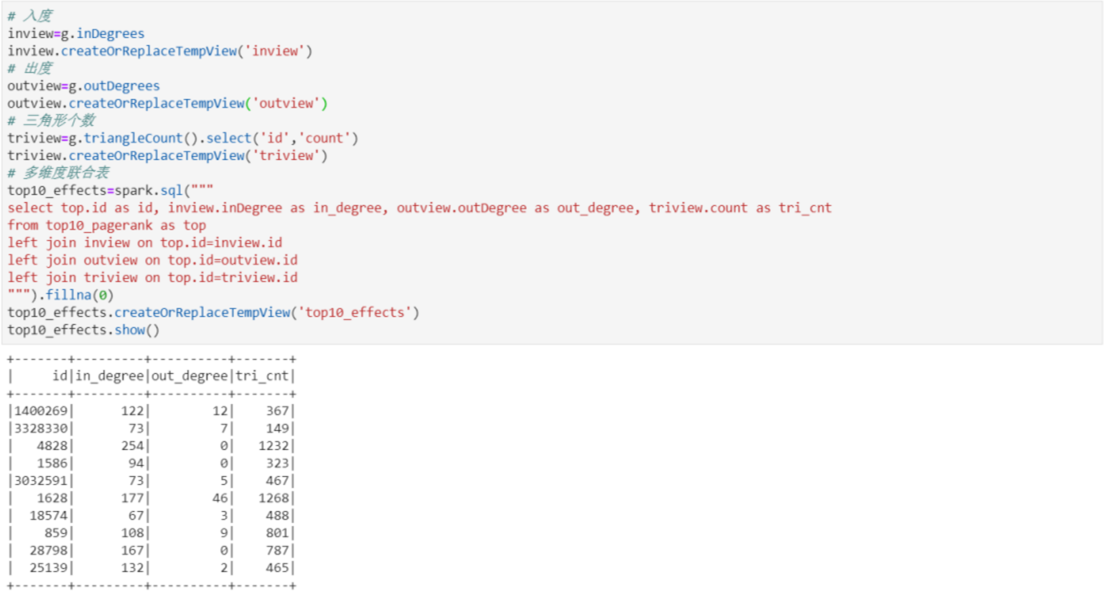
* 2: The original node and edge data files can be generated here for g ephi processing



```
# 获取pageRank前十
top10_pagerank=pr.vertices.sort(F.desc('pagerank')).limit(10)
top10_pagerank.createOrReplaceTempView('top10_pagerank')
top10_pagerank.show()
```

```
+-------+--------------+--------------------+-------------------+----+----+-------+----------+----------+-------+----------------+-----------+--------------------+------------------+
|     id|         login|             company|         created_at|type|fake|deleted|      long|       lat|country|           state|       city|            location|          pagerank|
+-------+--------------+--------------------+-------------------+----+----+-------+----------+----------+-------+----------------+-----------+--------------------+------------------+
|   4828|        fabpot|SensioLabs/Symfon...|2009-01-17 12:42:51| USR|   0|      0|  3.057256|  50.62925|     fr|            Nord|      Lille|               Lille| 76.95261925455124|
|   1628|      Ocramius|Marco Pivetta Sof...|2009-11-17 07:18:49| USR|   0|      0| 8.6821267|50.1109221|     de|       Darmstadt|  Frankfurt|   Frankfurt am Main|38.872596257878186|
|  25139|         nikic|                null|2010-03-04 20:22:25| USR|   0|      0| 13.404954|52.5200066|     de|          Berlin|     Berlin|     Berlin, Germany|28.382310086860475|
|  28798|       Seldaek|           Packagist|2010-01-16 17:28:47| USR|   0|      0|       0.0|       0.0|   null|            null|       null|Zürich, Zurich, S...| 28.06993967257823|
|   1586| weierophinney|   Zend Technologies|2008-09-23 15:49:25| USR|   0|      0|-96.728333|43.5473028|     us|Minnehaha County|Sioux Falls|Sioux Falls, SD, USA|27.885335077618365|
|3328330|    freekmurze|             @spatie|2010-11-16 12:38:15| USR|   0|      0| 4.4024643|51.2194475|     be|         Antwerp|    Antwerp|    Antwerp, Belgium| 26.32948798866407|
|  18574|    schmittjoh|    Scrutinizer GmbH|2010-02-04 18:41:50| USR|   0|      0|       0.0|       0.0|   null|            null|       null|                null|21.252655636349125|
|1400269|GrahamCampbell|  University of York|2013-03-21 22:18:49| USR|   0|      0|       0.0|       0.0|   null|            null|       null|  The United Kingdom| 20.91038968856148|
|3032591|     webmozart|                null|2013-10-17 08:36:38| USR|   0|      0|16.3738189|48.2081743|     at|          Vienna|     Vienna|     Vienna, Austria|20.394705159811924|
|    859|          stof|          @Incenteev|2010-10-14 11:56:42| USR|   0|      0| 2.3522219| 48.856614|     fr|           Paris|      Paris|               Paris|19.119035474741896|
+-------+--------------+--------------------+-------------------+----+----+-------+----------+----------+-------+----------------+-----------+--------------------+------------------+
```

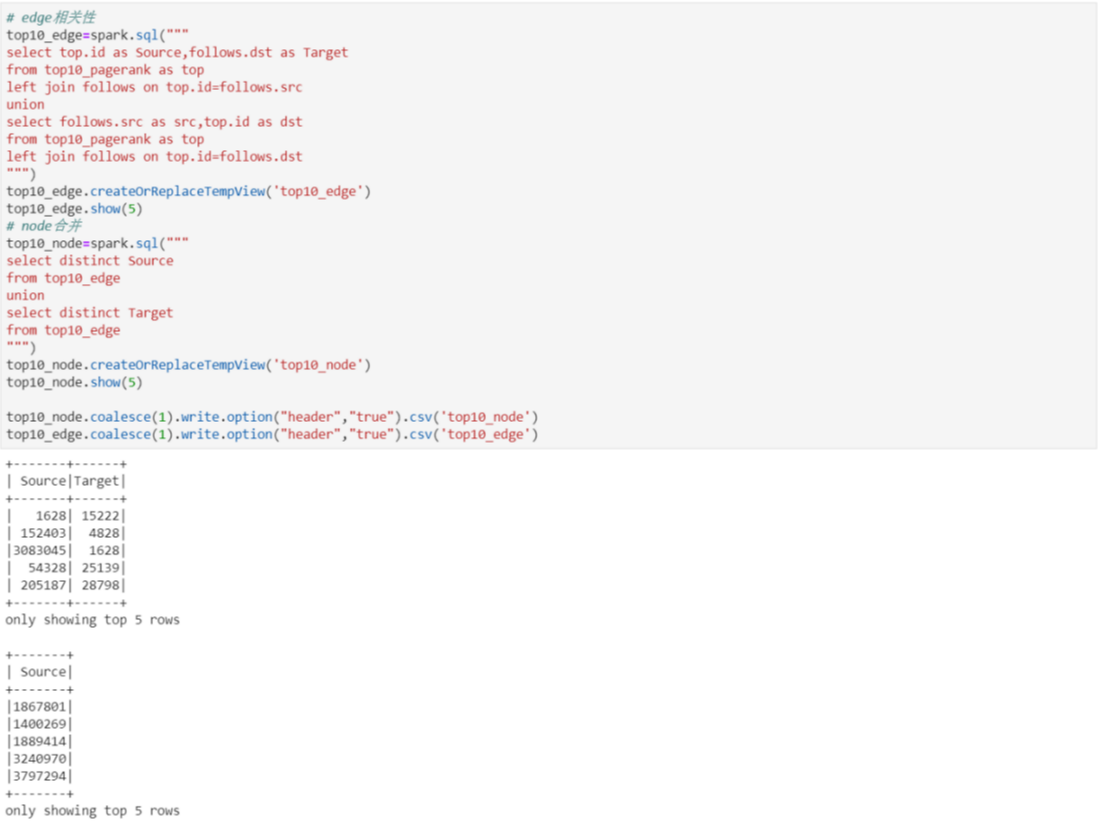
③ Obtain the top 10 information of PageRank



```
# 入度
inview=g.inDegrees
inview.createOrReplaceTempView('inview')
# 出度
outview=g.outDegrees
outview.createOrReplaceTempView('outview')
# 三角形个数
triview=g.triangleCount().select('id','count')
triview.createOrReplaceTempView('triview')
# 多维度联合表
top10_effects=spark.sql("""
select top.id as id, inview.inDegree as in_degree, outview.outDegree as out_degree, triview.count as tri_cnt
from top10_pagerank as top
left join inview on top.id=inview.id
left join outview on top.id=outview.id
left join triview on top.id=triview.id
""").fillna(0)
top10_effects.createOrReplaceTempView('top10_effects')
top10_effects.show()
```

```
+-------+---------+----------+-------+
|     id|in_degree|out_degree|tri_cnt|
+-------+---------+----------+-------+
|1400269|      122|        12|    367|
|3328330|       73|         7|    149|
|   4828|      254|         0|   1232|
|   1586|       94|         0|    323|
|3032591|       73|         5|    467|
|   1628|      177|        46|   1268|
|  18574|       67|         3|    488|
|    859|      108|         9|    801|
|  28798|      167|         0|    787|
|  25139|      132|         2|    465|
+-------+---------+----------+-------+
```

④ Obtain multi-dimensional data such as in-degree, out-degree, and number of triangles of the project

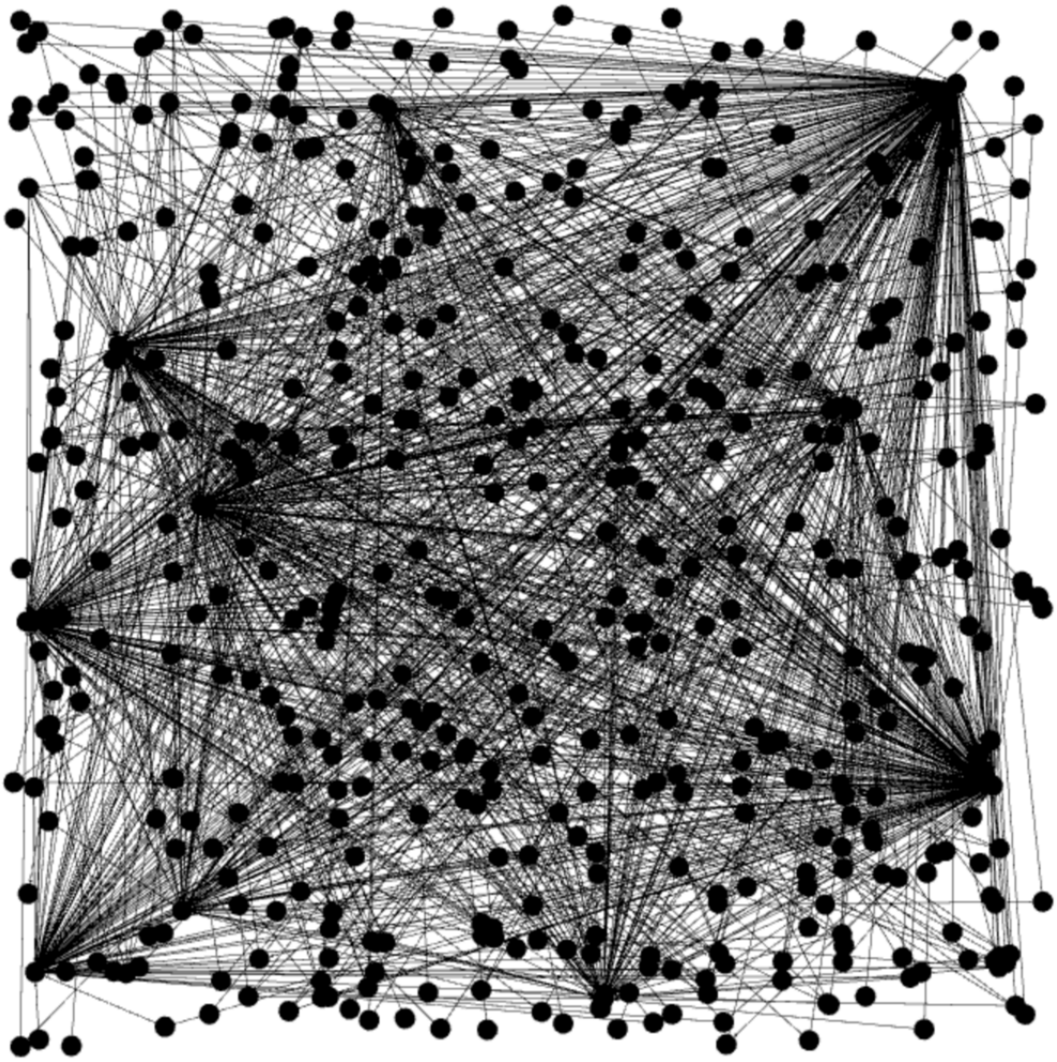


```
# edge相关性
top10_edge=spark.sql("""
select top.id as Source,follows.dst as Target
from top10_pagerank as top
left join follows on top.id=follows.src
union
select follows.src as src,top.id as dst
from top10_pagerank as top
left join follows on top.id=follows.dst
""")
top10_edge.createOrReplaceTempView('top10_edge')
top10_edge.show(5)
# node合并
top10_node=spark.sql("""
select distinct Source
from top10_edge
union
select distinct Target
from top10_edge
""")
top10_node.createOrReplaceTempView('top10_node')
top10_node.show(5)

top10_node.coalesce(1).write.option("header","true").csv('top10_node')
top10_edge.coalesce(1).write.option("header","true").csv('top10_edge')
```

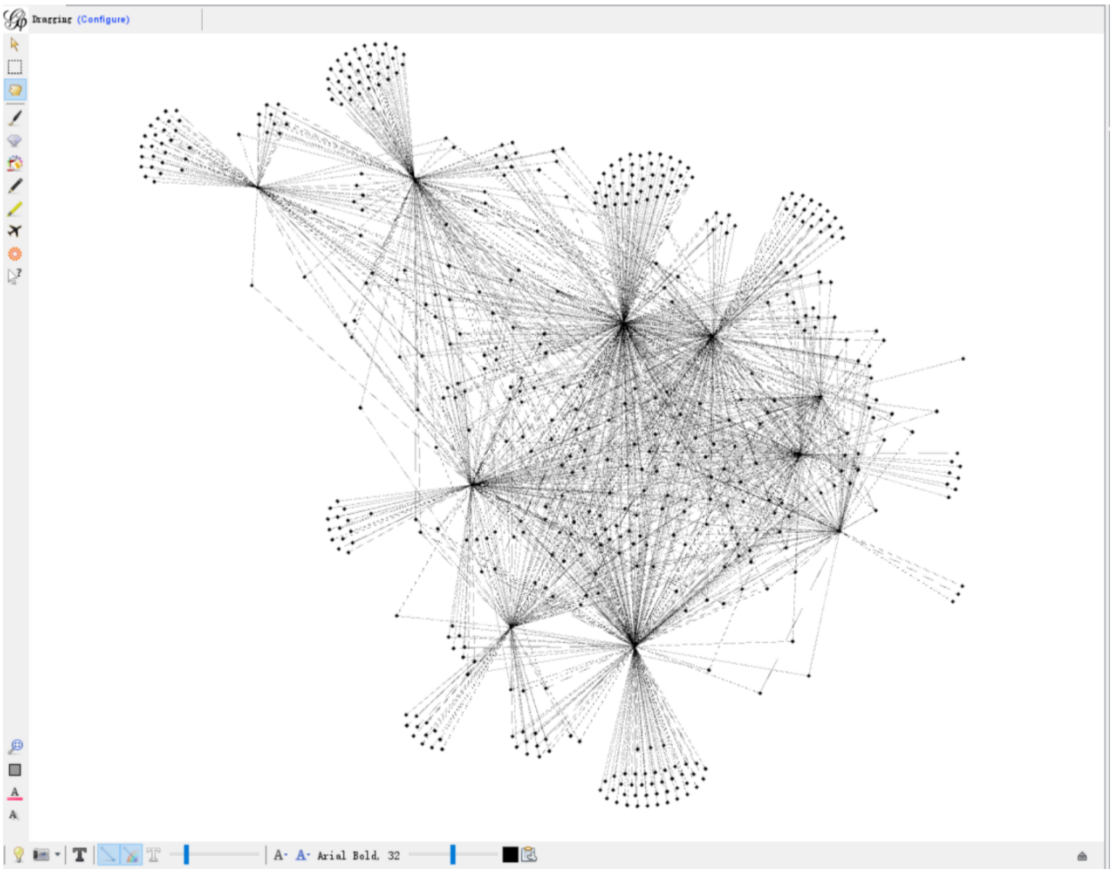
```
+-------+------+
| Source|Target|
+-------+------+
|   1628| 15222|
| 152403|  4828|
|3083045|  1628|
|  54328| 25139|
| 205187| 28798|
+-------+------+
only showing top 5 rows

+-------+
| Source|
+-------+
|1867801|
|1400269|
|1889414|
|3240970|
|3797294|
+-------+
only showing top 5 rows
```

as node and edge from the top 10 for Gephi processing

Gephi relational network generating the top ten data (as follows)



Relationship network diagram of top _10



The stretched top10 relationship network diagram

From the picture above, we can clearly see the top10 attention relationship network, and the top 10 is where the 10 obvious edges are concentrated .

【Experiment Summary】

In this class, I learned the method of identifying the developer's influence, consulted the information, learned and mastered the PageRank calculation method, the GraphFrames construction method and the related network index calculation interface, and analyzed it based on the data provided. At the beginning of the experiment, be careful to back up the original files first, because the original environment must be removed when configuring the environment for this experiment. When using Gephi to generate images, pay attention to the use of data formats, especially because the software only provides brief translations in multiple languages, so it is necessary to use the English version as much as possible. If you use Chinese, some data formats will make mistakes. Finally, after this experiment, I have enhanced my understanding of GraphFrame es , improved data analysis and application capabilities, and laid a solid foundation for future practice.

Through the learning of the H adoop experimental course, I have strengthened my understanding of big data and can effectively use cutting-edge technologies such as s park for data analysis. This course is newer, harder but also more challenging than most of our other courses, which I benefited from a lot!