Lab_4_python_3

December 1, 2022

[1]:
```
import findspark
from pyspark import SparkConf
from pyspark import SparkContext
from pyspark.sql import SparkSession

findspark.init()
spark = SparkContext.getOrCreate(SparkConf().setMaster("local[4]"))
spark = SparkSession(spark)
```

[22]:
```
from graphframes import *
from pyspark.sql.types import *
from pyspark.sql import functions as F
```

[4]:
```
v = spark.read.csv("/home/spark/lab04/task3/social-nodes.csv", header=True)
e = spark.read.csv("/home/spark/lab04/task3/social-relationships.csv",␣
↪header=True)
g = GraphFrame(v, e)
```

[5]:
```
results = g.pageRank(resetProbability=0.15, maxIter=20)
results.vertices.sort("pagerank", ascending=False).show()
```

```
+-------+------------------+
|     id|          pagerank|
+-------+------------------+
|   Doug| 2.2865372087512252|
|   Mark| 2.1424484186137263|
|  Alice|  1.520330830262095|
|Michael| 0.7274429252585624|
|Bridget| 0.7274429252585624|
|Charles| 0.5213852310709753|
|    Amy| 0.5097143486157744|
|  David|0.36655842368870073|
|  James| 0.1981396884803788|
+-------+------------------+
```

[6]:
```
results = g.pageRank(resetProbability=0.15, tol=0.01)
results.vertices.sort("pagerank", ascending=False).show
```

[6]:
```
<bound method DataFrame.show of DataFrame[id: string, pagerank: double]>
```

[7]:
```
total_degree = g.degrees
in_degree = g.inDegrees
out_degree = g.outDegrees
(total_degree.join(in_degree, "id", how="left")
  .join(out_degree, "id", how="left")
  .fillna(0)
  .sort("inDegree", ascending=False)
  .show())
```

```
+-------+------+--------+---------+
|     id|degree|inDegree|outDegree|
+-------+------+--------+---------+
|   Doug|     6|       5|        1|
|  Alice|     7|       3|        4|
|Bridget|     5|       2|        3|
|Michael|     5|       2|        3|
|    Amy|     1|       1|        0|
|   Mark|     3|       1|        2|
|  David|     2|       1|        1|
|Charles|     2|       1|        1|
|  James|     1|       0|        1|
+-------+------+--------+---------+
```

[19]:
```
import matplotlib
import matplotlib.pyplot as plt
from pyspark.sql import functions as F
plt.style.use('fivethirtyeight')
```

[9]:
```
nodes = spark.read.csv("/home/spark/lab04/task4/airports.csv", header=False)
```

[14]:
```
nodes.columns
```

[14]:
```
['_c0',
 '_c1',
 '_c2',
 '_c3',
 '_c4',
 '_c5',
 '_c6',
 '_c7',
 '_c8',
```

```
'_c9',
'_c10',
'_c11',
'_c12',
'_c13']
```

[15]:
```
cleaned_nodes = (nodes.select("_c1", "_c3", "_c4", "_c6", "_c7")
                 .filter("_c3 = 'United States'")
                 .withColumnRenamed("_c1", "name")
                 .withColumnRenamed("_c4", "id")
                 .withColumnRenamed("_c6", "latitude")
                 .withColumnRenamed("_c7", "longitude")
                 .drop("_c3"))
```

[16]:
```
cleaned_nodes = cleaned_nodes[cleaned_nodes["id"] != "\\N"]
```

[21]:
```
relationships = spark.read.csv("/home/spark/lab04/task4/188591317_T_ONTIME.
↪csv", header=True)
```

[23]:
```
cleaned_relationships = (relationships
                         .select("ORIGIN", "DEST", "FL_DATE", "DEP_DELAY",␣
↪"ARR_DELAY",
                                 "DISTANCE", "TAIL_NUM", "FL_NUM",␣
↪"CRS_DEP_TIME",
                                 "CRS_ARR_TIME", "UNIQUE_CARRIER")
                         .withColumnRenamed("ORIGIN", "src")
                         .withColumnRenamed("DEST", "dst")
                         .withColumnRenamed("DEP_DELAY", "deptDelay")
                         .withColumnRenamed("ARR_DELAY", "arrDelay")
                         .withColumnRenamed("TAIL_NUM", "tailNumber")
                         .withColumnRenamed("FL_NUM", "flightNumber")
                         .withColumnRenamed("FL_DATE", "date")
                         .withColumnRenamed("CRS_DEP_TIME", "time")
                         .withColumnRenamed("CRS_ARR_TIME", "arrivalTime")
                         .withColumnRenamed("DISTANCE", "distance")
                         .withColumnRenamed("UNIQUE_CARRIER", "airline")
                         .withColumn("deptDelay", F.col("deptDelay").
↪cast(FloatType()))
                         .withColumn("arrDelay", F.col("arrDelay").
↪cast(FloatType()))
                         .withColumn("time", F.col("time").cast(IntegerType()))
                         .withColumn("arrivalTime", F.col("arrivalTime").
↪cast(IntegerType()))
                         )
```

[24]:
```
g = GraphFrame(cleaned_nodes, cleaned_relationships)
```

[25]:
```
airlines_reference = (spark.read.csv("/home/spark/lab04/task4/airlines.csv")
      .select("_c1", "_c3")
      .withColumnRenamed("_c1", "name")
      .withColumnRenamed("_c3", "code"))
```

[26]:
```
airlines_reference = airlines_reference[airlines_reference["code"] != "null"]
```

[27]:
```
df = spark.read.option("multiline", "true").json("/home/spark/lab04/task4/
 ↪airlines.json")
dummyDf = spark.createDataFrame([("test", "test")], ["code", "name"])
```

[28]:
```
for code in df.schema.fieldNames():
    tempDf = (df.withColumn("code", F.lit(code))
              .withColumn("name", df[code]))
    tdf = tempDf.select("code", "name")
    dummyDf = dummyDf.union(tdf)
```

[29]:
```
g.vertices.count()
```

[29]:
```
1333
```

[30]:
```
g.edges.count()
```

[30]:
```
616529
```

[31]:
```
g.edges.groupBy().max("deptDelay").show()
```

```
+--------------+
|max(deptDelay)|
+--------------+
|        1632.0|
+--------------+
```

[32]:
```
airports_degree = g.outDegrees.withColumnRenamed("id", "oId")
```

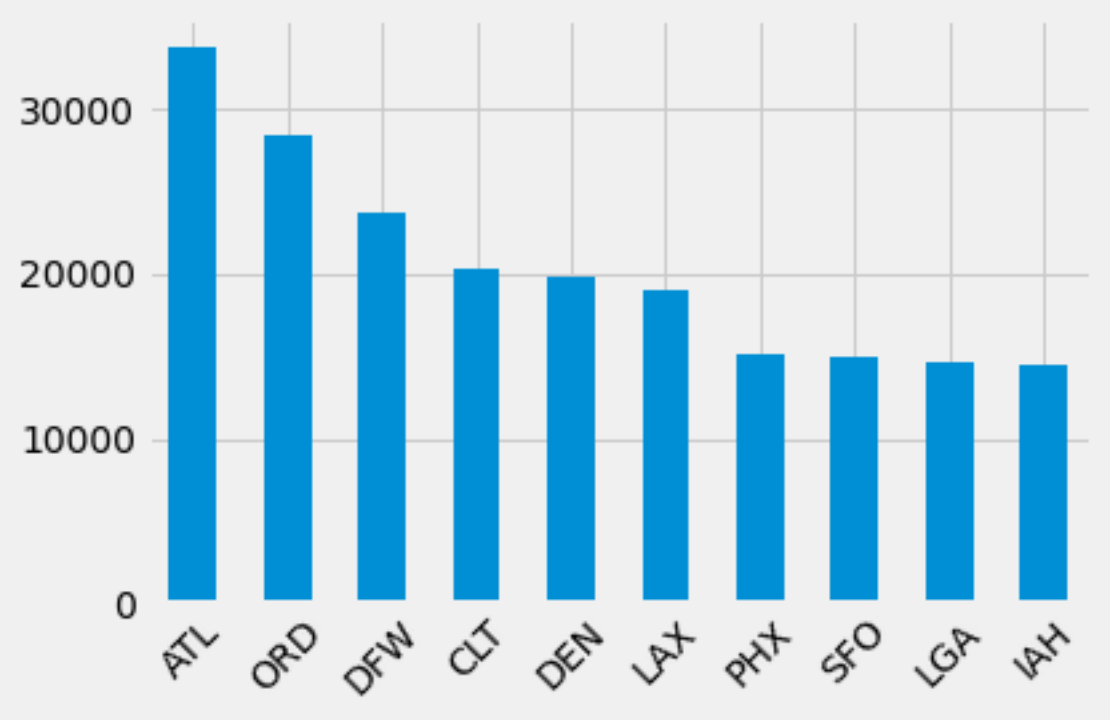
[33]:
```
full_airports_degree = (airports_degree
                        .join(g.vertices, airports_degree.oId == g.vertices.id)
                        .sort("outDegree", ascending=False)
                        .select("id", "name", "outDegree"))
```

[34]:
```
full_airports_degree.show(n=10, truncate=False)
```

```
+---+-----------------------------------------------+---------+
|id |name                                           |outDegree|
+---+-----------------------------------------------+---------+
|ATL|Hartsfield Jackson Atlanta International Airport|33837    |
```

```
|ORD|Chicago O'Hare International Airport            |28338    |
|DFW|Dallas Fort Worth International Airport         |23765    |
|CLT|Charlotte Douglas International Airport         |20251    |
|DEN|Denver International Airport                    |19836    |
|LAX|Los Angeles International Airport               |19059    |
|PHX|Phoenix Sky Harbor International Airport        |15103    |
|SFO|San Francisco International Airport             |14934    |
|LGA|La Guardia Airport                              |14709    |
|IAH|George Bush Intercontinental Houston Airport    |14407    |
+---+-----------------------------------------------+---------+
only showing top 10 rows
```

[35]:
```
plt.style.use('fivethirtyeight')

ax = (full_airports_degree
      .toPandas()
      .head(10)
      .plot(kind='bar', x='id', y='outDegree', legend=None))

ax.xaxis.set_label_text("")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



[36]:
```
delayed_flights = (g.edges
                   .filter("src = 'ORD' and deptDelay > 0")
                   .groupBy("dst")
                   .agg(F.avg("deptDelay"), F.count("deptDelay"))
                   .withColumn("averageDelay", F.round(F.col("avg(deptDelay)"),␣
 ↪2))
                   .withColumn("numberOfDelays", F.col("count(deptDelay)")))
```

[37]:
```
(delayed_flights
 .join(g.vertices, delayed_flights.dst == g.vertices.id)
 .sort(F.desc("averageDelay"))
 .select("dst", "name", "averageDelay", "numberOfDelays")
 .show(n=10, truncate=False))
```

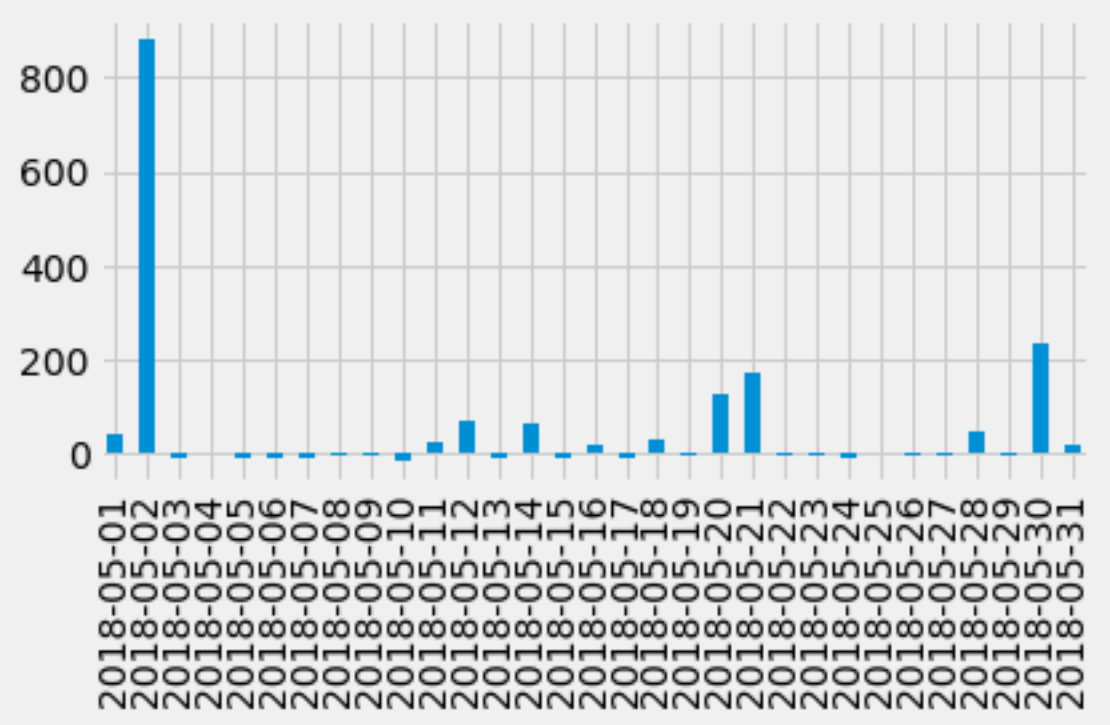
```
+---+-----------------------------------------+------------+--------------+
|dst|name                                     |averageDelay|numberOfDelays|
+---+-----------------------------------------+------------+--------------+
|CKB|North Central West Virginia Airport      |145.08      |12            |
|OGG|Kahului Airport                          |119.67      |9             |
|MQT|Sawyer International Airport             |114.75      |12            |
|MOB|Mobile Regional Airport                  |102.2       |10            |
|TTN|Trenton Mercer Airport                   |101.18      |17            |
|AVL|Asheville Regional Airport               |98.5        |28            |
|ISP|Long Island Mac Arthur Airport           |94.08       |13            |
|ANC|Ted Stevens Anchorage International Airport|83.74       |23            |
|BTV|Burlington International Airport         |83.2        |25            |
|CMX|Houghton County Memorial Airport         |79.18       |17            |
+---+-----------------------------------------+------------+--------------+
only showing top 10 rows
```

[38]:
```
from_expr = 'id = "ORD"'
to_expr = 'id = "CKB"'
ord_to_ckb = g.bfs(from_expr, to_expr)
```

[39]:
```
ord_to_ckb = ord_to_ckb.select(
  F.col("e0.date"),
  F.col("e0.time"),
  F.col("e0.flightNumber"),
  F.col("e0.deptDelay"))
```

[40]:
```
ax = (ord_to_ckb
      .sort("date")
      .toPandas()
      .plot(kind='bar', x='date', y='deptDelay', legend=None))

ax.xaxis.set_label_text("")
```

```
plt.tight_layout()
plt.show()
```



[52]:
```
motifs = (g.find("(a)-[ab]->(b); (b)-[bc]->(c)")
          .filter("""(b.id = 'SFO') and
                  (ab.date = '2018-05-11' and bc.date = '2018-05-11') and
                  (ab.arrDelay > 30 or bc.deptDelay > 30) and
                  (ab.flightNumber = bc.flightNumber) and
                  (ab.airline = bc.airline) and
                  (ab.time < bc.time)"""))
```

[53]:
```
def sum_dist(dist1, dist2):
    return sum([value for value in [dist1, dist2] if value is not None])
```

[54]:
```
sum_dist_udf = F.udf(sum_dist, FloatType())
```

[55]:
```
result = (motifs.withColumn("delta", motifs.bc.deptDelay - motifs.ab.arrDelay)
          .select("ab", "bc", "delta")
          .sort("delta", ascending=False))
```

[56]:
```
result.select(
    F.col("ab.src").alias("a1"),
    F.col("ab.time").alias("a1DeptTime"),
    F.col("ab.arrDelay"),
```

```
    F.col("ab.dst").alias("a2"),
    F.col("bc.time").alias("a2DeptTime"),
    F.col("bc.deptDelay"),
    F.col("bc.dst").alias("a3"),
    F.col("ab.airline"),
    F.col("ab.flightNumber"),
    F.col("delta")
).show()
```

```
+---+---------+--------+---+---------+---------+---+-------+-----------+-----
+
| a1|a1DeptTime|arrDelay| a2|a2DeptTime|deptDelay|
a3|airline|flightNumber|delta|
+---+---------+--------+---+---------+---------+---+-------+-----------+-----
+
|PDX|      1130|   -18.0|SFO|      1350|    178.0|BUR|     WN|
1454|196.0|
|ACV|      1755|    -9.0|SFO|      2235|     64.0|RDM|     OO|        5700|
73.0|
|BWI|       700|    -3.0|SFO|      1125|     49.0|IAD|     UA|         753|
52.0|
|ATL|       740|    40.0|SFO|      1110|     77.0|SAN|     UA|        1900|
37.0|
|BUR|      1405|    25.0|SFO|      1600|     39.0|PDX|     WN|         157|
14.0|
|DTW|       835|    34.0|SFO|      1135|     44.0|DTW|     DL|         745|
10.0|
|DEN|      1830|    25.0|SFO|      2045|     33.0|BUR|     WN|        1783|
8.0|
|PDX|      1855|   119.0|SFO|      2120|    117.0|DEN|     WN|        5789|
-2.0|
|BUR|      2025|    31.0|SFO|      2230|     11.0|PHX|     WN|
1585|-20.0|
+---+---------+--------+---+---------+---------+---+-------+-----------+-----
+
```

[57]:
```
result = g.pageRank(resetProbability=0.15, maxIter=20)
(result.vertices
 .sort("pagerank", ascending=False)
 .withColumn("pagerank", F.round(F.col("pagerank"), 2))
 .show(truncate=False, n=100))
```

```
+-----------------------------------------------------------+---+-------------
----+------------------+--------+
|name                                                       |id |latitude
|longitude          |pagerank|
+-----------------------------------------------------------+---+-------------
```

```
----+------------------+--------+
|Hartsfield Jackson Atlanta International Airport
|ATL|33.63669967651367 |-84.4281005859375  |45.17   |
|Chicago O'Hare International Airport                          |ORD|41.97859955
|-87.90480042       |40.35   |
|Dallas Fort Worth International Airport
|DFW|32.89680099487305 |-97.03800201416016 |37.18   |
|Denver International Airport
|DEN|39.861698150635   |-104.672996521     |28.93   |
|Charlotte Douglas International Airport
|CLT|35.2140007019043  |-80.94309997558594 |27.65   |
|Los Angeles International Airport                             |LAX|33.94250107
|-118.4079971       |22.29   |
|Minneapolis-St Paul International/Wold-Chamberlain Airport
|MSP|44.881999969499994|-93.22180175780001 |22.0    |
|Detroit Metropolitan Wayne County Airport
|DTW|42.212398529052734|-83.35340118408203 |20.42   |
|Phoenix Sky Harbor International Airport
|PHX|33.43429946899414 |-112.01200103759766|19.09   |
|George Bush Intercontinental Houston Airport
|IAH|29.984399795532227|-95.34140014648438 |19.03   |
|San Francisco International Airport
|SFO|37.61899948120117 |-122.375           |18.87   |
|McCarran International Airport                                |LAS|36.08010101
|-115.1520004       |17.23   |
|Seattle Tacoma International Airport
|SEA|47.44900131225586 |-122.30899810791016|17.18   |
|La Guardia Airport                                            |LGA|40.77719879
|-73.87259674       |17.1    |
|Salt Lake City International Airport
|SLC|40.78839874267578 |-111.97799682617188|16.24   |
|General Edward Lawrence Logan International Airport           |BOS|42.36429977
|-71.00520325       |14.61   |
|Newark Liberty International Airport
|EWR|40.692501068115234|-74.168701171875   |14.43   |
|John F Kennedy International Airport                          |JFK|40.63980103
|-73.77890015       |13.35   |
|Orlando International Airport
|MCO|28.429399490356445|-81.30899810791016 |13.16   |
|Ronald Reagan Washington National Airport                     |DCA|38.8521
|-77.037697         |12.96   |
|Philadelphia International Airport
|PHL|39.87189865112305 |-75.24109649658203 |12.4    |
|Baltimore/Washington International Thurgood Marshall Airport|BWI|39.17539978
|-76.66829681       |10.88   |
|Fort Lauderdale Hollywood International Airport
|FLL|26.072599411010742|-80.15270233154297 |9.52    |
|San Diego International Airport                               |SAN|32.7336006165
```

```
|-117.190002441     |9.05    |
|Chicago Midway International Airport
|MDW|41.7859992980957  |-87.75240325927734 |8.76    |
|Miami International Airport
|MIA|25.79319953918457 |-80.29060363769531 |8.28    |
|Washington Dulles International Airport                    |IAD|38.94449997
|-77.45580292       |7.35    |
|Nashville International Airport
|BNA|36.1245002746582  |-86.6781997680664  |7.26    |
|Tampa International Airport
|TPA|27.975500106811523|-82.533203125      |7.23    |
|Dallas Love Field
|DAL|32.84709930419922 |-96.85179901123047 |7.09    |
|Lambert St Louis International Airport
|STL|38.74869918823242 |-90.37000274658203 |6.74    |
|Austin Bergstrom International Airport
|AUS|30.194499969482422|-97.6698989868164  |6.68    |
|Portland International Airport                             |PDX|45.58869934
|-122.5979996       |6.58    |
|Raleigh Durham International Airport
|RDU|35.877601623535156|-78.7874984741211  |6.25    |
|Ted Stevens Anchorage International Airport
|ANC|61.174400329589844|-149.99600219726562|6.03    |
|William P Hobby Airport                                    |HOU|29.64539909
|-95.27890015       |5.9     |
|Kansas City International Airport                          |MCI|39.2976
|-94.713898         |5.84    |
|Louis Armstrong New Orleans International Airport
|MSY|29.99340057373047 |-90.25800323486328 |5.83    |
|Norman Y. Mineta San Jose International Airport
|SJC|37.36259841918945 |-121.92900085449219|5.68    |
|Cincinnati Northern Kentucky International Airport         |CVG|39.0488014221
|-84.6678009033     |5.55    |
|Metropolitan Oakland International Airport
|OAK|37.72129821777344 |-122.22100067138672|5.48    |
|Honolulu International Airport
|HNL|21.318700790405273|-157.9219970703125 |5.37    |
|Indianapolis International Airport                         |IND|39.7173
|-86.294403         |5.29    |
|Pittsburgh International Airport                           |PIT|40.49150085
|-80.23290253       |5.11    |
|Cleveland Hopkins International Airport                    |CLE|41.4117012024
|-81.8498001099     |4.99    |
|Sacramento International Airport
|SMF|38.69540023803711 |-121.59100341796875|4.98    |
|Port Columbus International Airport
|CMH|39.99800109863281 |-82.89189910888672 |4.57    |
|John Wayne Airport-Orange County Airport                   |SNA|33.67570114
```

```
|-117.8679962       |4.48    |
|San Antonio International Airport
|SAT|29.533700942993164|-98.46980285644531 |4.09    |
|General Mitchell International Airport
|MKE|42.947200775146484|-87.89659881591797 |3.63    |
|Jacksonville International Airport
|JAX|30.49410057067871 |-81.68789672851562 |3.51    |
|Orlando Sanford International Airport
|SFB|28.777599334716797|-81.23750305175781 |3.18    |
|Bradley International Airport                                |BDL|41.9388999939
|-72.68319702149999 |3.04    |
|Eppley Airfield
|OMA|41.303199768066406|-95.89409637451172 |2.99    |
|Southwest Florida International Airport
|RSW|26.53619956970215 |-81.75520324707031 |2.96    |
|Bob Hope Airport
|BUR|34.20069885253906 |-118.35900115966797|2.91    |
|Charleston Air Force Base-International Airport              |CHS|32.89860153
|-80.04049683       |2.82    |
|Buffalo Niagara International Airport                        |BUF|42.94049835
|-78.73220062       |2.82    |
|Kahului Airport
|OGG|20.89859962463379 |-156.42999267578125|2.8     |
|Will Rogers World Airport
|OKC|35.39310073852539 |-97.60070037841797 |2.8     |
|Memphis International Airport
|MEM|35.04240036010742 |-89.97669982910156 |2.76    |
|Palm Beach International Airport
|PBI|26.68320083618164 |-80.09559631347656 |2.75    |
|Albuquerque International Sunport Airport
|ABQ|35.040199279785156|-106.60900115966797|2.74    |
|Louisville International Standiford Field                    |SDF|38.1744
|-85.736            |2.74    |
|Richmond International Airport
|RIC|37.50519943237305 |-77.3197021484375  |2.62    |
|Norfolk International Airport
|ORF|36.89459991455078 |-76.20120239257812 |2.6     |
|Ontario International Airport
|ONT|34.055999755859375|-117.60099792480469|2.49    |
|Phoenix-Mesa-Gateway Airport                                 |AZA|33.30780029
|-111.6549988       |2.47    |
|Boise Air Terminal/Gowen field                               |BOI|43.56439972
|-116.2229996       |2.4     |
|St Petersburg Clearwater International Airport               |PIE|27.91020012
|-82.68740082       |2.37    |
|Juneau International Airport
|JNU|58.35499954223633 |-134.5760040283203 |2.29    |
|Tucson International Airport
```

|TUS|32.1161003112793 |-110.94100189208984|2.25 |
|Myrtle Beach International Airport |MYR|33.6796989441
|-78.9282989502 |2.23 |
|Theodore Francis Green State Airport
|PVD|41.732601165771484|-71.42040252685547 |2.2 |
|McGhee Tyson Airport |TYS|35.81100082
|-83.9940033 |2.18 |
|Des Moines International Airport
|DSM|41.534000396728516|-93.66310119628906 |2.17 |
|Tulsa International Airport
|TUL|36.19839859008789 |-95.88809967041016 |2.16 |
|El Paso International Airport |ELP|31.80719948
|-106.3779984 |2.16 |
|Birmingham-Shuttlesworth International Airport |BHM|33.56290054
|-86.75350189 |2.16 |
|Reno Tahoe International Airport
|RNO|39.49909973144531 |-119.76799774169922|2.12 |
|Gerald R. Ford International Airport |GRR|42.88079834
|-85.52279663 |2.1 |
|Long Beach /Daugherty Field/ Airport |LGB|33.81769943
|-118.1520004 |2.09 |
|Savannah Hilton Head International Airport |SAV|32.12760162
|-81.20210266 |2.0 |
|Bill & Hillary Clinton National Airport/Adams Field
|LIT|34.729400634799994|-92.2242965698 |1.83 |
|James M Cox Dayton International Airport
|DAY|39.902400970458984|-84.21939849853516 |1.83 |
|Kona International At Keahole Airport
|KOA|19.738800048828125|-156.04600524902344|1.82 |
|Northwest Arkansas Regional Airport |XNA|36.281898
|-94.306801 |1.75 |
|Dane County Regional Truax Field
|MSN|43.13990020751953 |-89.3375015258789 |1.75 |
|Greater Rochester International Airport
|ROC|43.118900299072266|-77.67240142822266 |1.75 |
|Syracuse Hancock International Airport
|SYR|43.11119842529297 |-76.1063003540039 |1.74 |
|Greenville Spartanburg International Airport |GSP|34.8956985474
|-82.2189025879 |1.73 |
|Lihue Airport
|LIH|21.97599983215332 |-159.33900451660156|1.72 |
|Spokane International Airport
|GEG|47.61989974975586 |-117.53399658203125|1.7 |
|Piedmont Triad International Airport
|GSO|36.097801208496094|-79.93730163574219 |1.68 |
|Albany International Airport
|ALB|42.74829864501953 |-73.80169677734375 |1.67 |
|Fairbanks International Airport |FAI|64.81510162

```
|-147.8560028       |1.65    |
|Pensacola Regional Airport
|PNS|30.473400115967   |-87.186599731445   |1.59    |
|Blue Grass Airport
|LEX|38.0364990234375  |-84.60590362548828 |1.57    |
|Fresno Yosemite International Airport
|FAT|36.77619934082031 |-119.71800231933594|1.52    |
|Charlotte County Airport                                        |PGD|26.92020035
|-81.9905014        |1.51    |
+-----------------------------------------------------------------+---+--------------
----+------------------+--------+
only showing top 100 rows
```

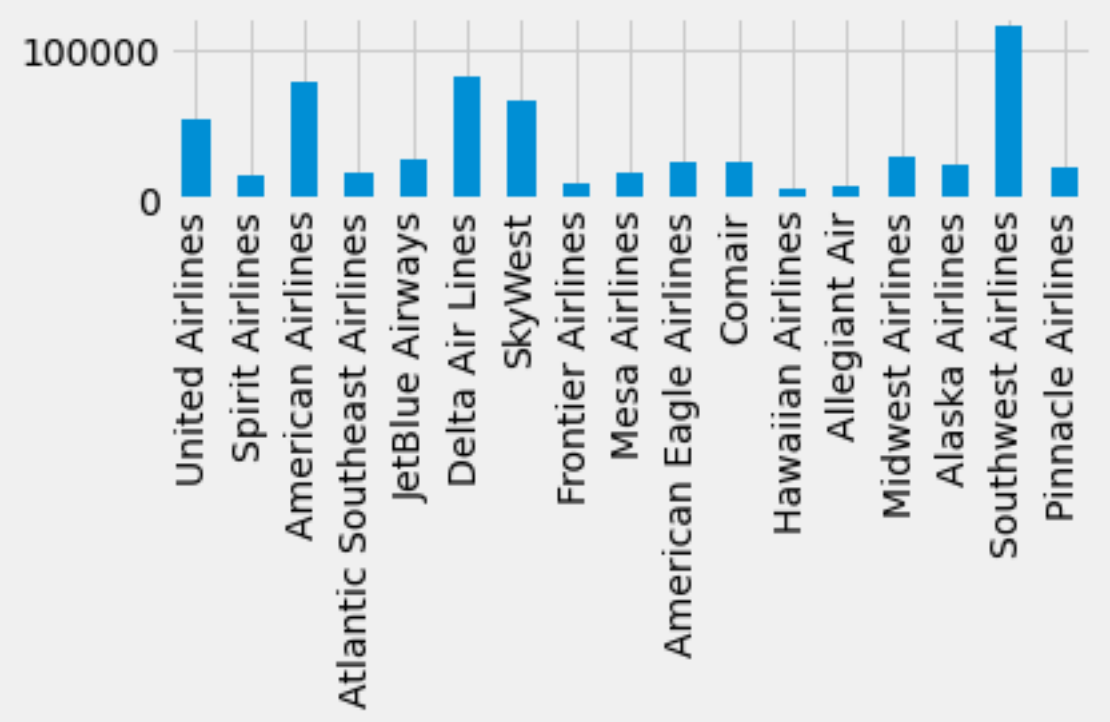
[58]:
```
triangles = g.triangleCount().cache()
pagerank = g.pageRank(resetProbability=0.15, maxIter=20).cache()

(triangles.select(F.col("id").alias("tId"), "count")
 .join(pagerank.vertices, F.col("tId") == F.col("id"))
 .select("id", "name", "pagerank", "count")
 .sort("count", ascending=False)
 .withColumn("pagerank", F.round(F.col("pagerank"), 2))
 .show(truncate=False))
```

```
+---+-----------------------------------------------------------------+--------+-----
+
|id |name
|pagerank|count|
+---+-----------------------------------------------------------------+--------+-----
+
|ATL|Hartsfield Jackson Atlanta International Airport            |45.17   |1783
|
|DEN|Denver International Airport                                |28.93   |1706
|
|ORD|Chicago O'Hare International Airport                        |40.35   |1669
|
|CLT|Charlotte Douglas International Airport                     |27.65   |1595
|
|DFW|Dallas Fort Worth International Airport                     |37.18   |1595
|
|LAS|McCarran International Airport                              |17.23   |1448
|
|MSP|Minneapolis-St Paul International/Wold-Chamberlain Airport  |22.0    |1412
|
|DTW|Detroit Metropolitan Wayne County Airport                   |20.42   |1401
|
|PHX|Phoenix Sky Harbor International Airport                    |19.09   |1333
|
```

```
|IAH|George Bush Intercontinental Houston Airport                |19.03   |1266
|
|DCA|Ronald Reagan Washington National Airport                    |12.96   |1194
|
|EWR|Newark Liberty International Airport                         |14.43   |1191
|
|LAX|Los Angeles International Airport                            |22.29   |1191
|
|MCO|Orlando International Airport                                |13.16   |1172
|
|PHL|Philadelphia International Airport                           |12.4    |1109
|
|SEA|Seattle Tacoma International Airport                         |17.18   |1072
|
|BWI|Baltimore/Washington International Thurgood Marshall Airport|10.88   |1065
|
|BOS|General Edward Lawrence Logan International Airport          |14.61   |1061
|
|AUS|Austin Bergstrom International Airport                       |6.68    |1056
|
|FLL|Fort Lauderdale Hollywood International Airport              |9.52    |1025
|
+---+-------------------------------------------------------------+--------+-----
+
only showing top 20 rows
```

[59]:
```
airlines = (g.edges
  .groupBy("airline")
  .agg(F.count("airline").alias("flights"))
  .sort("flights", ascending=False))
full_name_airlines = (airlines_reference
  .join(airlines, airlines.airline == airlines_reference.code)
  .select("code", "name", "flights"))
```

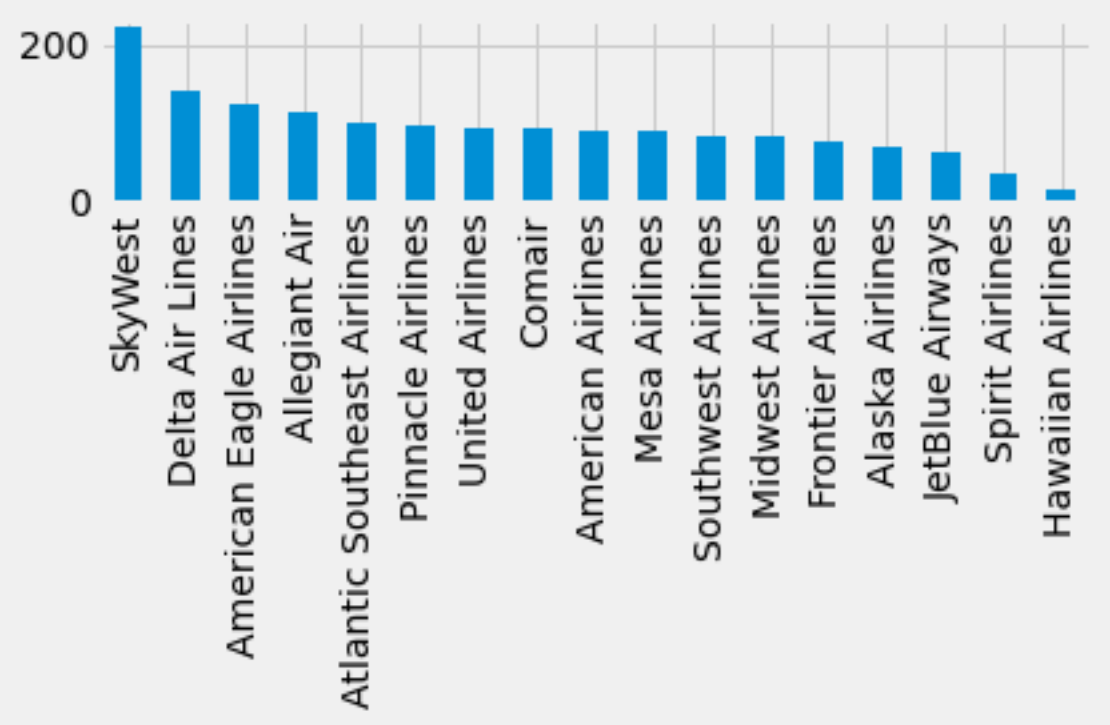
[60]:
```
ax = (full_name_airlines.toPandas()
  .plot(kind='bar', x='name', y='flights', legend=None))
ax.xaxis.set_label_text("")
plt.tight_layout()
plt.show()
```



[61]:
```
def find_scc_components(g, airline):
   # Create a subgraph containing only flights on the provided airline
   airline_relationships = g.edges[g.edges.airline == airline]
   airline_graph = GraphFrame(g.vertices, airline_relationships)
   # Calculate the Strongly Connected Components
   scc = airline_graph.stronglyConnectedComponents(maxIter=10)
   # Find the size of the biggest component and return that
   return (scc
     .groupBy("component")
     .agg(F.count("id").alias("size"))
     .sort("size", ascending=False)
     .take(1)[0]["size"])
```

[62]:
```
# Calculate the largest strongly connected component for each airline
airline_scc = [(airline, find_scc_components(g, airline))
for airline in airlines.toPandas()["airline"].tolist()]
airline_scc_df = spark.createDataFrame(airline_scc, ['id', 'sccCount'])
# Join the SCC DataFrame with the airlines DataFrame so that we can show
# the number of flights an airline has alongside the number of
# airports reachable in its biggest component
airline_reach = (airline_scc_df
.join(full_name_airlines, full_name_airlines.code == airline_scc_df.id)
.select("code", "name", "flights", "sccCount")
.sort("sccCount", ascending=False))
```

[63]:
```
ax = (airline_reach.toPandas()
.plot(kind='bar', x='name', y='sccCount', legend=None))
ax.xaxis.set_label_text("")
plt.tight_layout()
plt.show()
```



[64]:
```
airline_relationships = g.edges.filter("airline = 'DL'")
airline_graph = GraphFrame(g.vertices, airline_relationships)

clusters = airline_graph.labelPropagation(maxIter=10)
(clusters
 .sort("label")
 .groupby("label")
 .agg(F.collect_list("id").alias("airports"),
      F.count("id").alias("count"))
 .sort("count", ascending=False)
 .show(truncate=70, n=10))
```

```
+-------------+----------------------------------------------------------------
-----+-----+
|        label|
airports|count|
+-------------+----------------------------------------------------------------
-----+-----+
```

```
|1606317768706|[IND, ORF, ATW, RIC, TRI, XNA, ECP, AVL, JAX, SYR, BHM, GSO, MEM,
C…|   89|
|1219770712067|[GEG, SLC, DTW, LAS, SEA, BOS, MSN, SNA, JFK, TVC, LIH, JAC, FLL,
M…|   53|
|  17179869187|
[RHV]|    1|
|  25769803777|
[CWT]|    1|
|  25769803776|
[CDW]|    1|
|            1|
[CNU]|    1|
|  25769803778|
[DRT]|    1|
|  25769803779|
[FOK]|    1|
|            0|
[BGM]|    1|
|            2|
[DAW]|    1|
+-------------+--------------------------------------------------------------
-----+-----+
only showing top 10 rows
```

[65]:
```
all_flights = g.degrees.withColumnRenamed("id", "aId")
```

[66]:
```
(clusters
.filter("label=1606317768706")
.join(all_flights, all_flights.aId == clusters.id)
 .sort("degree", ascending=False)
.select("id", "name", "degree")
.show(truncate=False))
```

```
+---+-----------------------------------------------------------+------+
|id |name                                                       |degree|
+---+-----------------------------------------------------------+------+
|DFW|Dallas Fort Worth International Airport                    |47514 |
|CLT|Charlotte Douglas International Airport                    |40495 |
|IAH|George Bush Intercontinental Houston Airport               |28814 |
|EWR|Newark Liberty International Airport                       |25131 |
|PHL|Philadelphia International Airport                         |20804 |
|BWI|Baltimore/Washington International Thurgood Marshall Airport|18989 |
|MDW|Chicago Midway International Airport                       |15178 |
|BNA|Nashville International Airport                            |12455 |
|DAL|Dallas Love Field                                          |12084 |
|IAD|Washington Dulles International Airport                    |11566 |
|STL|Lambert St Louis International Airport                     |11439 |
```

```
|HOU|William P Hobby Airport                                    |9742  |
|IND|Indianapolis International Airport                         |8543  |
|PIT|Pittsburgh International Airport                           |8410  |
|CLE|Cleveland Hopkins International Airport                    |8238  |
|CMH|Port Columbus International Airport                        |7640  |
|SAT|San Antonio International Airport                          |6532  |
|JAX|Jacksonville International Airport                         |5495  |
|BDL|Bradley International Airport                              |4866  |
|RSW|Southwest Florida International Airport                    |4569  |
+---+-----------------------------------------------------------+------+
only showing top 20 rows
```

[ ]: