ISOM 673 Social Network Analytics
Final Paper
Professor Demetrius Lewis
Meng Cheng, Carl Xi, Jie Zhu

Skynet

"The Wright Brothers created the single greatest cultural force since the invention of writing. The airplane became the first World Wide Web, bringing people, languages, ideas, and values together." - Bill Gates

"Flight DL215 to New York will begin boarding in 5 minutes." Carl snapped back to reality as the lady on the PA chimed to life. It is a chilly February morning at London Heathrow International Airport. The weather is what you'd expect, a sea of grey fog, not a single sunray in sight. As Carl quietly observed the people eagerly lining up to board the plane, he thought back to his Social Network Analytics class with Professor Demetrius. Noticing the diversity of the people lining up, he wondered if he can create a network by using flight routes and airports as proxies for edges and nodes respectively. Carl thought back to all the discussions on centrality measures, eigenvector, …. And wondered what interesting results can he drive from analyzing such a network.

Indeed, air travel has some really interesting properties that deserve a closer look. With over 4 billion people traveling via some 22,000 routes around the world per year, it has become the single most important way for goods and people to move around. Through this paper, we want to analyze the international air travel network in order to drive interesting insights through observing node, group or network-level effects. On the node level, we want to find airports that have outstanding properties, such as the most popular or important airports (the two might not be the same!), or airports that have the highest betweenness. On the group level, we want to find

cliques or transfer effects that may affect travel patterns. On the network level, we want to look at general trends that may have propagated through our network, as well as how the formation & stability of ties in the evolution of our network. Once we look at all three levels, we also want to run further analysis with external data. For example, it would be interesting to conduct a regression analysis using each node's network measures and their socioeconomic factors.

Dataset Description

In order to answer our questions at all these three levels, we use datasets from OpenFlights.org to conduct social network analysis. This website provides us with two useful databases: Route Database and Airport Database. The Route Database has data available for two time points, June 2014 and November 2019, while the Airport Database was updated January, 2017.

Route Database contains information of routes operated by commercial airlines across the world. Following are variables in the Route Database:

- **Airline** **2-letter (IATA) or 3-letter (ICAO) code of the airline.**
- **Airline ID** **Unique OpenFlights identifier for airlines.**
- **Source airport** **3-letter (IATA) or 4-letter (ICAO) code of the source airport.**
- **Source airport ID** **Unique OpenFlights identifier for source airport.**
- **Destination airport** **3-letter (IATA) or 4-letter (ICAO) code of the destination airport.**
- **Destination airport ID** **Unique OpenFlights identifier for destination airport.**
- **Codeshare** **"Y" if this flight is a codeshare, empty otherwise.**
- **Stops** **Number of stops on this flight ("0" for direct).**
- **Equipment** **3-letter codes for plane type(s) generally used on this flight.**

Airport Database contains information of airports and their city & country of affiliation. Following are variables in the Airport Database:

- **Airport ID** **Unique OpenFlights identifier for this airport.**
- **Name** **Name of airport. May or may not contain the City name.**

- **City** **Main city served by airport. May be spelled differently from Name.**
- **Country** **Country or territory where the airport is located.**
- **IATA** **3-letter IATA code. Null if not assigned/unknown.**
- **ICAO** **4-letter ICAO code.**
- **Latitude** **Decimal degrees. Negative is South, positive is North.**
- **Longitude** **Decimal degrees. Negative is West, positive is East.**
- **Altitude** **In feet.**
- **Timezone** **Hours offset from UTC. Fractional hours are expressed as decimals.**

Figure 1. Initial Plot of Routes (2014) Using Tableau

Figure 2. Initial Plot of Airports (2014) Using Tableau

After merging these two databases on *Airport ID*, we find 406 NAs in *Source City* and 416 NAs in *Destination City*. This is because some small airports are not listed in the Airport Database. Since these smaller ones are useful in calculating other airports' centrality measures, we decide to keep routes with NAs for now. The primary variables for social network analysis are *Source Airport, Source City, Source Country, Destination Airport, Destination City,* and *Destination Country*.

With the unioned dataset, we are able to show airports and their number of unique routes. Each circle in the following Tableau plot represents an airport, while the size of the circle demonstrates how many unique routes the airport has. Here, we can see that major international hub airports are all marked by big circles, which means the measure we choose, i.e. unique routes, can effectively reflect the importance of airports.

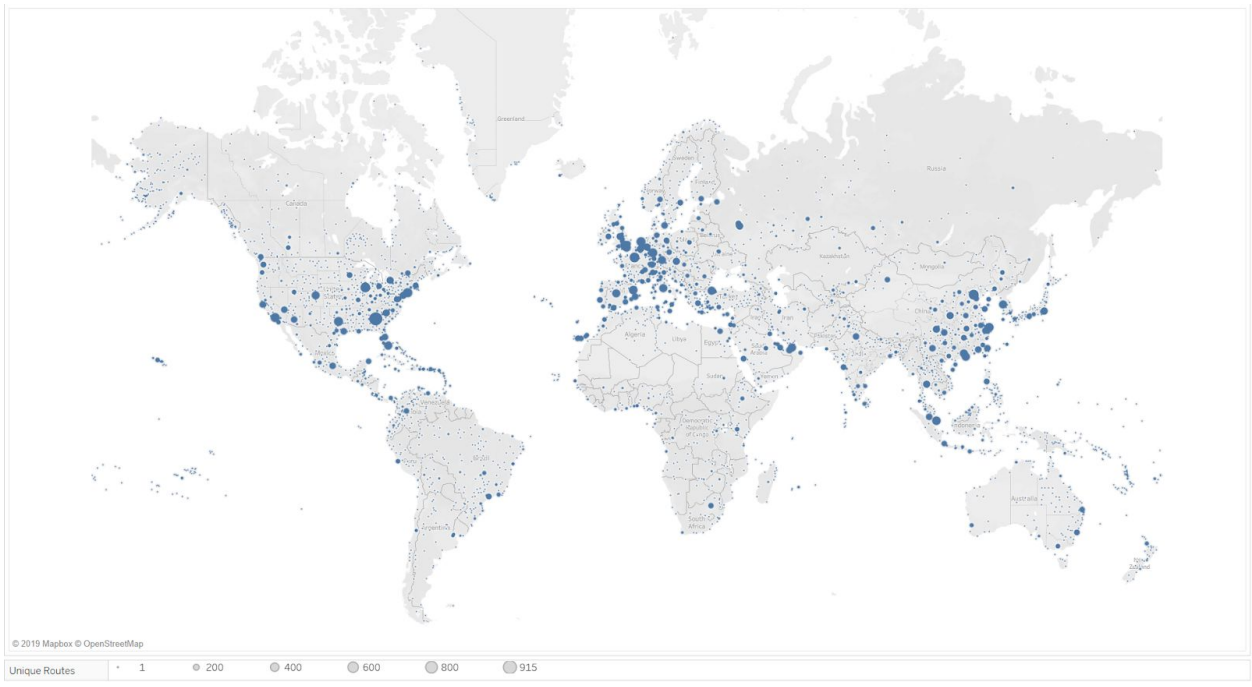


Figure 3. Airports and Number of Unique Routes (2014) Using Tableau

Exploratory Analysis

With a proper dataset in hand, we are now able to conduct some exploratory analyses using social network measures like betweenness, closeness, coreness, Eigenvector centrality, and hub centrality. Due to the geographical nature of the dataset, we primarily use Tableau to plot geospatial charts and find interesting facts.

In the following chart, we add a color layer to a city-level plot, so that cities with higher betweenness will have darker color. It appears that the top 3 cities in betweenness are Amsterdam, Tokyo, and Anchorage. This makes sense because:

(1) For Amsterdam, it is the biggest hub and the headquarter of KLM, one of the biggest airlines in Europe and one of the founders of the Sky Team. It also benefits from the freedoms of the air within the European Union. KLM connects Amsterdam with lots of small cities in Europe. Thus, Amsterdam has very high betweenness value.

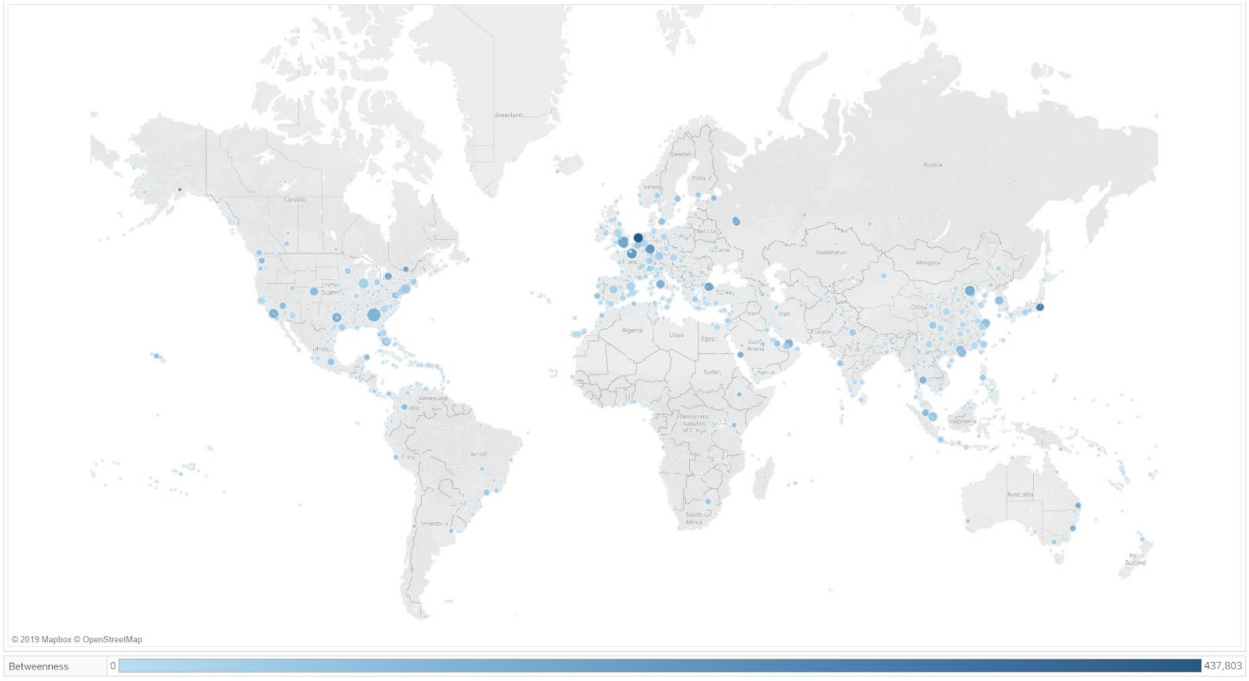


Figure 4. Cities and Betweenness (2014) Using Tableau

(2) For Tokyo, it is the biggest city of Japan and the headquarter city for both Japan Airlines and ANA. Also, Japan is highly developed in economy, and Japanese people have great demands for international travelling. Most airports in Japan and even in some other Asian countries have to connect to Tokyo in order to reach the rest of the world, which boosts up Tokyo's betweenness value.

(3) For Anchorage, the logic is similar. The only difference between Anchorage and the other two is that Anchorage is much smaller in population and in the size of the economy. However, Alaska is a state with lots of hills and mountains, which has lots of small airports that can only connect to other countries and states via Anchorage, the only major airport in Alaska.

Following are the Tableau charts with other social network measures. Some of these charts have similar colors across the map. This is because the measure used in the chart doesn't separate cities as much as other measures.

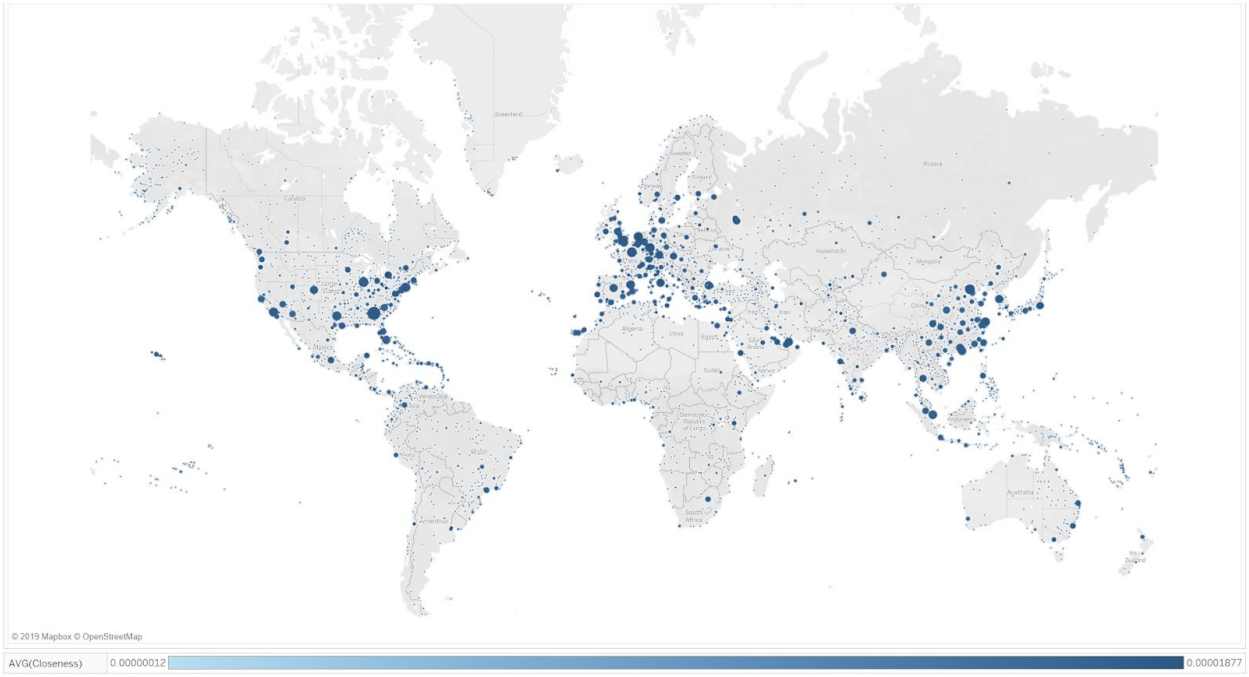


Figure 5. Cities and Closeness (2014) Using Tableau

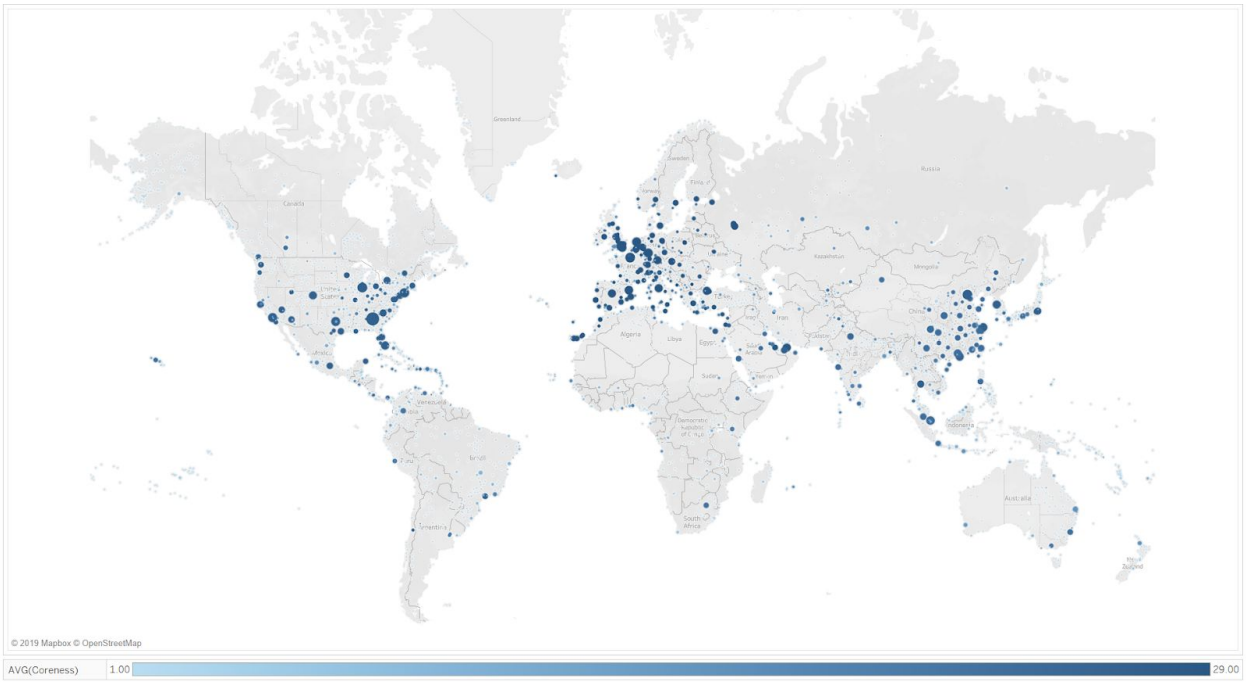


Figure 6. Cities and Coreness (2014) Using Tableau

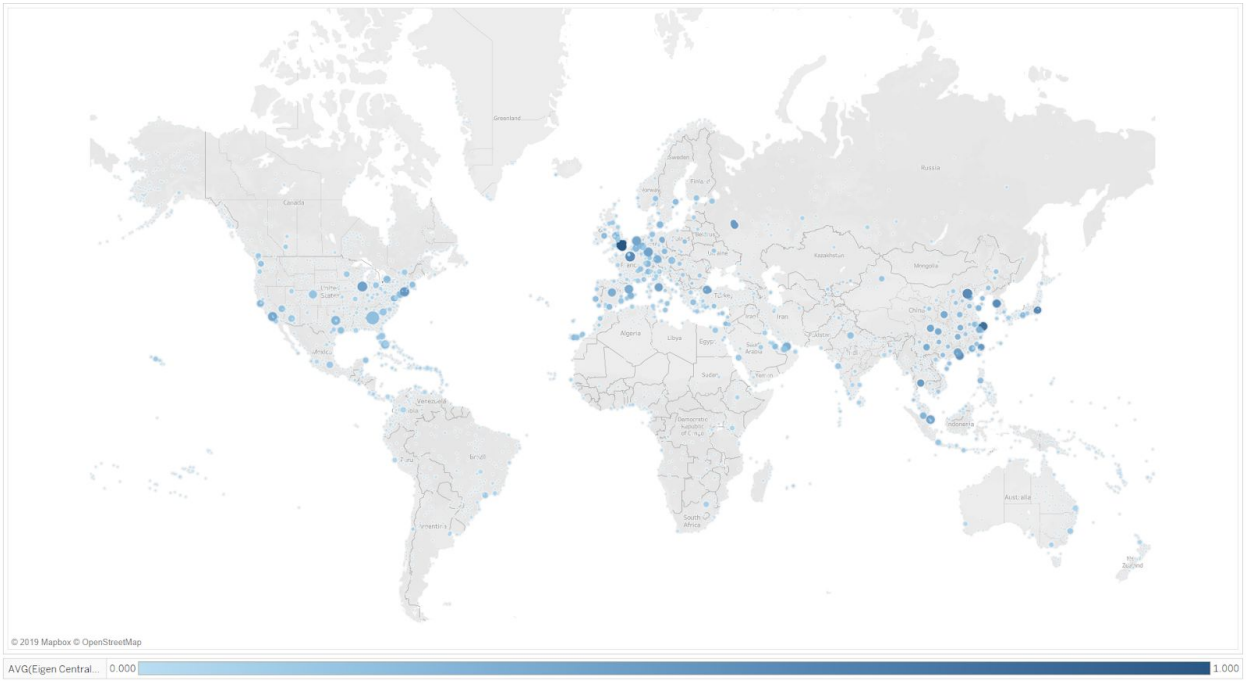


Figure 7. Cities and Eigen Centrality (2014) Using Tableau

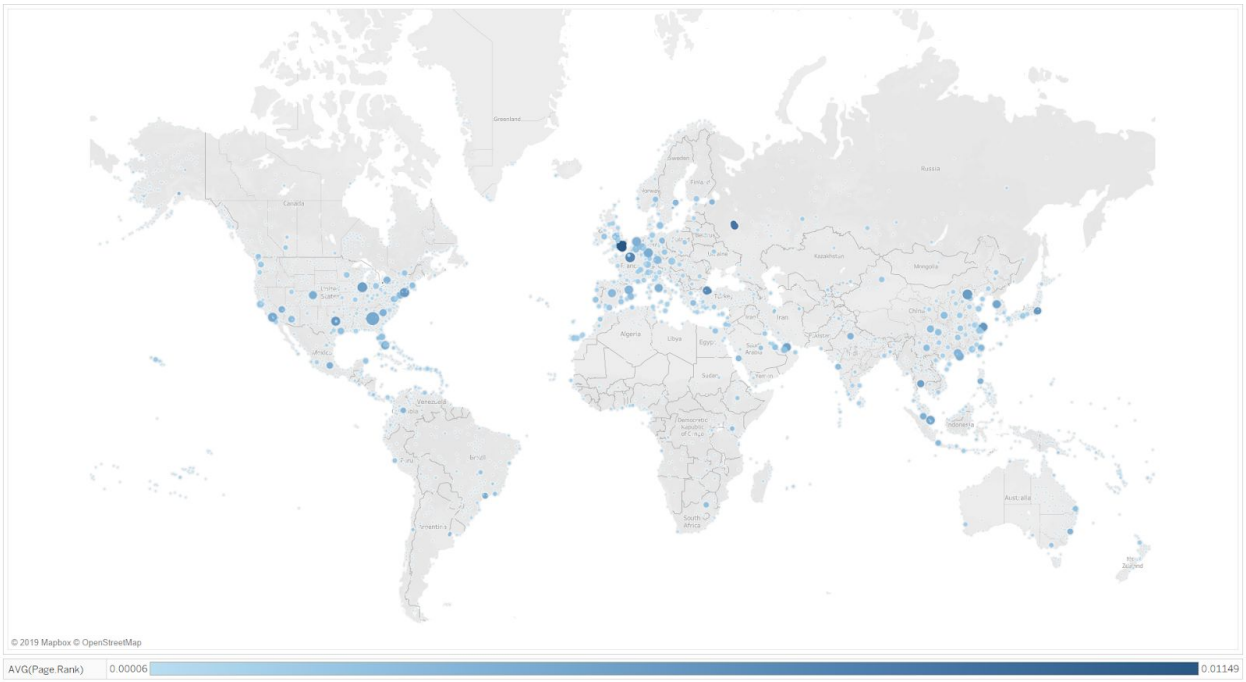


Figure 8. Cities and Page Rank (2014) Using Tableau

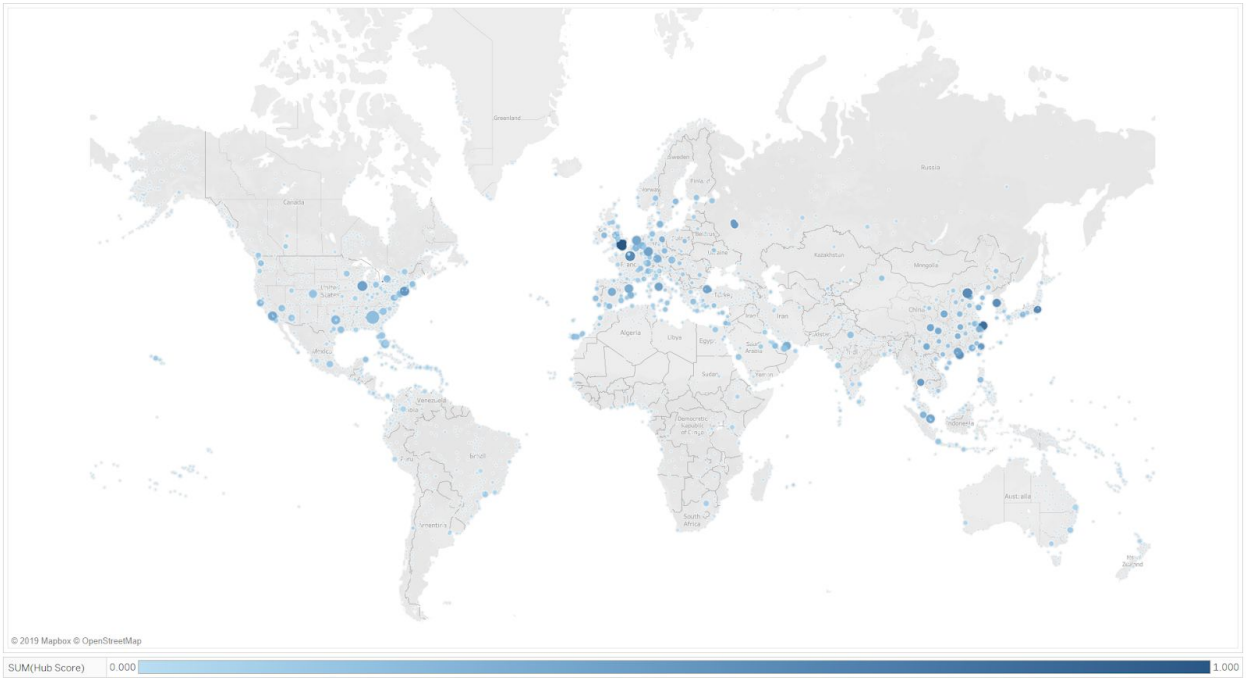
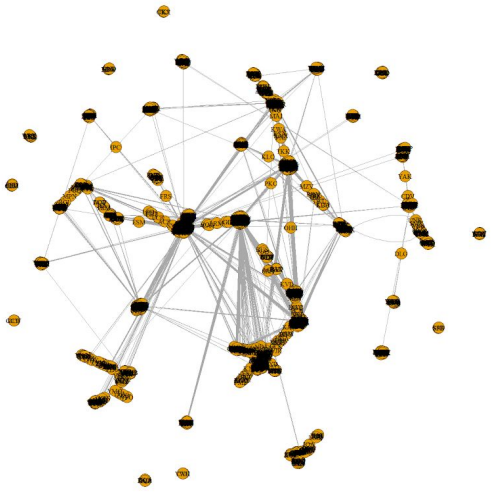


Figure 9. Cities and Hub Score (2014) Using Tableau

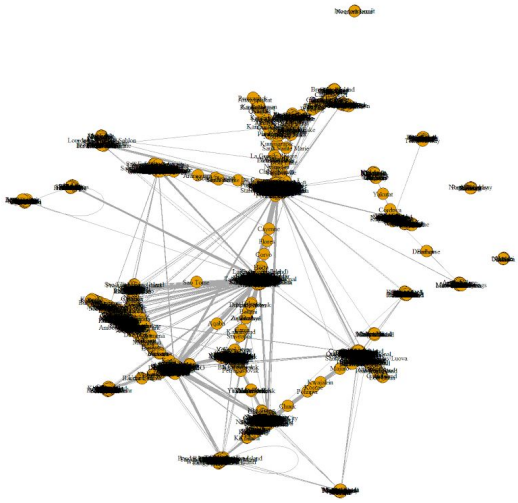
Analysis Methods Discussion

Firstly, to better show the relationship between each node and the hierarchical relationship, we visualize the full network in airport, city and country level with igraph in r. For country level, as we are only interested in international flights, we remove the edges represents for domestic flights and replot the network graph.

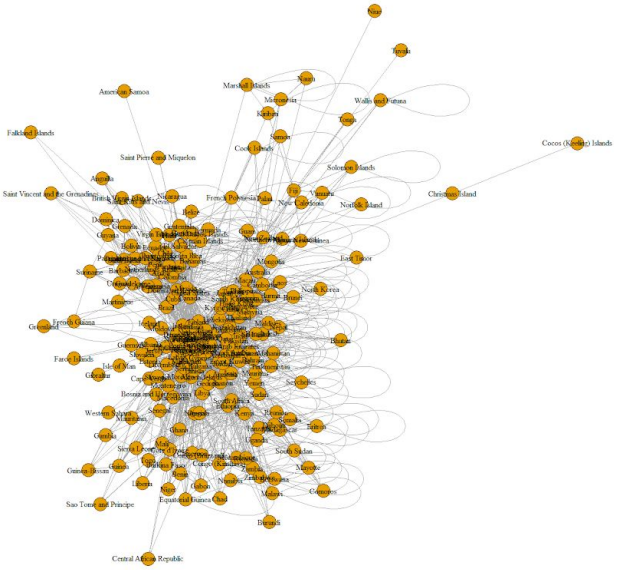
Full Raw Network



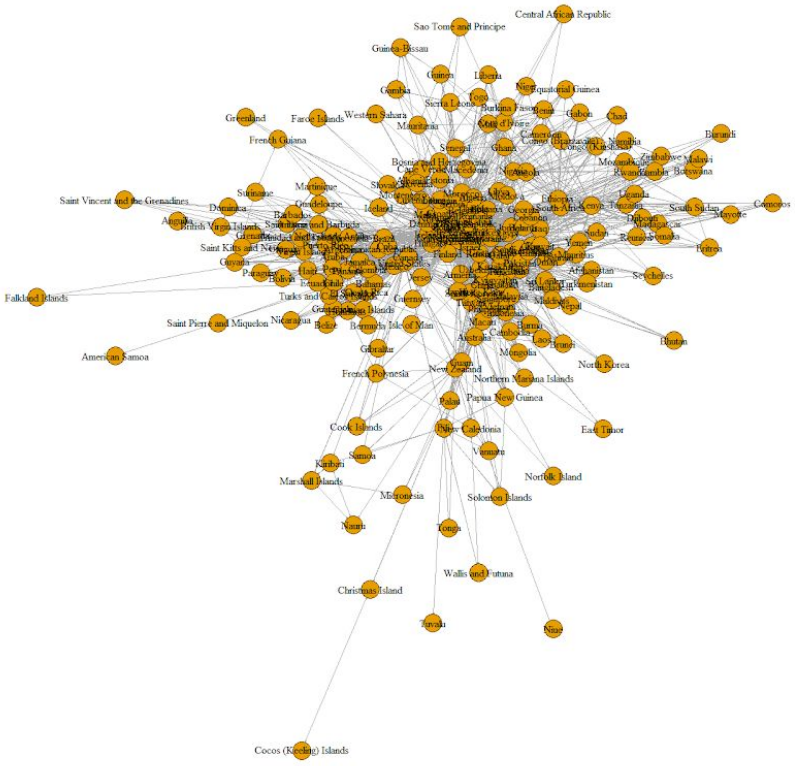
Cities Network



Countries Network



International Only Countries Network



Among all four networks of different levels, there are indeed clusterings exist. Take the international countries network as an example, Latin American countries are in the middle left; African countries are in the upper area; European and Asian countries are in the center playing roles as bridges between other continents (where we observe high betweenness in Asian and Eurpoean countries later). Also, North American countries such as the United States and Canada act specifically as bridges between Latin American countries and European and Asian countries.

Result Illustration

In order to find out if the centrality of a certain node is related to its socioeconomic characteristics, we combine below data from World Bank Database for both year 2014 and 2019 with our country level centrality measurements.

- **1"Educational attainment, at least a Bachelor's or equivalent, population 25+, total (%) (cumulative)"**
- **2 "Educational attainment, at least completed lower secondary, population 25+, total (%) (cumulative)"**
- **3 "Educational attainment, at least completed primary, population 25+ years, total (%) (cumulative)"**
- **4 "Employment to population ratio, 15+, total (%) (national estimate)"**
- **5 "GDP (current US$)"**
- **6 "GDP per capita (current US$)"**
- **7 "Individuals using the Internet (% of population)"**
- **8 "Population density (people per sq. km of land area)"**
- **9 "Population, male (% of total population)"**
- **10 "Population, total"**
- **11 "Trade in services (% of GDP)"**
- **12 "Urban population (% of total population)"**

As variable 5 is highly correlated to variable 6 and 10, we decide to get rid of it. We also choose to get rid of educational variable 1,2,3 due to missing data for small countries.

We ran generalized linear model (GLM) between each centrality measures and all available socioeconomic variables (centrality measure as the y variable). We choose glm model as it is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. Model results are shown below:

```
Call:
glm(formula = strength.2014 ~ ., data = countrywhole2014[, c(2,
    13, 15:21)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-914.17  -270.41  -107.25    87.78  2950.47

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.812e+03  1.065e+03   1.702  0.09192 .
`2014-4`    -1.061e+01  5.869e+00  -1.808  0.07372 .
`2014-6`     1.342e-02  4.509e-03   2.977  0.00367 **
`2014-7`     3.487e+00  4.133e+00   0.844  0.40085
`2014-8`     7.167e-02  8.393e-02   0.854  0.39524
`2014-9`    -2.222e+01  2.086e+01  -1.065  0.28948
`2014-10`    1.930e-06  4.363e-07   4.425 2.51e-05 ***
`2014-11`   -5.589e+00  1.915e+00  -2.919  0.00436 **
`2014-12`   -8.524e-01  4.322e+00  -0.197  0.84409
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 375978.2)

    Null deviance: 58596229  on 106  degrees of freedom
Residual deviance: 36845861  on  98  degrees of freedom
  (70 observations deleted due to missingness)
AIC: 1687.8

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = coreness.2014 ~ ., data = countrywhole2014[, c(3,
    13, 15:21)])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-10.3987   -3.1412    0.6348    3.4009    9.5129

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.038e+01  7.791e+00   2.616  0.01032 *
`2014-4`    -1.161e-01  4.294e-02  -2.703  0.00810 **
`2014-6`     5.024e-05  3.299e-05   1.523  0.13103
`2014-7`     1.228e-01  3.024e-02   4.060 9.9e-05 ***
`2014-8`     6.953e-04  6.141e-04   1.132  0.26032
`2014-9`    -1.063e-01  1.526e-01  -0.696  0.48783
`2014-10`    9.017e-09  3.192e-09   2.825  0.00573 **
`2014-11`   -1.834e-02  1.401e-02  -1.309  0.19368
`2014-12`   -1.014e-02  3.163e-02  -0.321  0.74905
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 20.1259)

    Null deviance: 4042.9  on 106  degrees of freedom
Residual deviance: 1972.3  on  98  degrees of freedom
  (70 observations deleted due to missingness)
AIC: 635.47

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = page.rank.2014 ~ ., data = countrywhole2014[, c(4,
    13, 15:21)])

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.013389  -0.003352  -0.001466   0.001031   0.035300

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.356e-02  1.367e-02   1.724  0.08784 .
`2014-4`    -1.179e-04  7.532e-05  -1.565  0.12085
`2014-6`     1.873e-07  5.788e-08   3.236  0.00165 **
`2014-7`     2.468e-05  5.304e-05   0.465  0.64271
`2014-8`     1.023e-06  1.077e-06   0.950  0.34443
`2014-9`    -2.934e-04  2.678e-04  -1.096  0.27586
`2014-10`    2.757e-11  5.599e-12   4.924 3.43e-06 ***
`2014-11`   -7.870e-05  2.458e-05  -3.202  0.00184 **
`2014-12`    5.862e-06  5.548e-05   0.106  0.91606
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 6.193053e-05)

    Null deviance: 0.0100621  on 106  degrees of freedom
Residual deviance: 0.0060692  on  98  degrees of freedom
  (70 observations deleted due to missingness)
AIC: -722.52

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = betweenness.2014 ~ ., data = countrywhole2014[,
    c(5, 13, 15:21)])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1054.40   -241.79    -86.88    103.58   2033.05

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.542e+00  7.721e+02  -0.005  0.99635
`2014-4`    -2.853e+00  4.256e+00  -0.670  0.50419
`2014-6`     9.729e-03  3.270e-03   2.976  0.00368 **
`2014-7`     1.653e+00  2.997e+00   0.552  0.58253
`2014-8`     4.560e-02  6.086e-02   0.749  0.45550
`2014-9`     4.744e+00  1.513e+01   0.314  0.75451
`2014-10`    1.755e-07  3.163e-07   0.555  0.58032
`2014-11`   -4.578e+00  1.388e+00  -3.297  0.00136 **
`2014-12`    1.368e+00  3.134e+00   0.437  0.66337
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 197667.5)

    Null deviance: 27379621  on 106  degrees of freedom
Residual deviance: 19371415  on  98  degrees of freedom
  (70 observations deleted due to missingness)
AIC: 1619

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = closeness.2014 ~ ., data = countrywhole2014[, c(6,
    13, 15:21)])

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.965e-04  -6.492e-05   6.267e-06   6.520e-05   1.832e-04

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.753e-04  1.699e-04   4.563 1.46e-05 ***
`2014-4`    -1.105e-06  9.365e-07  -1.180  0.24103
`2014-6`     1.870e-09  7.196e-10   2.598  0.01082 *
`2014-7`     1.787e-07  6.595e-07   0.271  0.78693
`2014-8`     1.103e-08  1.339e-08   0.824  0.41211
`2014-9`     2.350e-06  3.329e-06   0.706  0.48189
`2014-10`    8.038e-14  6.962e-14   1.155  0.25107
`2014-11`   -8.256e-07  3.055e-07  -2.702  0.00812 **
`2014-12`   -3.934e-07  6.897e-07  -0.570  0.56972
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 9.573361e-09)

    Null deviance: 1.1501e-06  on 106  degrees of freedom
Residual deviance: 9.3819e-07  on  98  degrees of freedom
  (70 observations deleted due to missingness)
AIC: -1661.4

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = eigen_centrality.2014 ~ ., data = countrywhole2014[,
    c(7, 13, 15:21)])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.21412  -0.07045  -0.02746   0.02248   0.82206

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.101e-01  2.641e-01   1.932   0.0563 .
`2014-4`    -3.077e-03  1.455e-03  -2.114   0.0371 *
`2014-6`     2.849e-06  1.118e-06   2.548   0.0124 *
`2014-7`     9.565e-04  1.025e-03   0.933   0.3530
`2014-8`     3.186e-06  2.081e-05   0.153   0.8786
`2014-9`    -6.196e-03  5.173e-03  -1.198   0.2340
`2014-10`    1.341e-10  1.082e-10   1.239   0.2182
`2014-11`   -1.074e-03  4.748e-04  -2.261   0.0260 *
`2014-12`   -4.872e-04  1.072e-03  -0.455   0.6504
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.02311933)

    Null deviance: 2.9998  on 106  degrees of freedom
Residual deviance: 2.2657  on  98  degrees of freedom
  (70 observations deleted due to missingness)
AIC: -88.827

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = authority ~ ., data = countrywhole2014[, c(8, 13,
    15:21)])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.21109  -0.07033  -0.02730   0.02236   0.82386

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.054e-01  2.606e-01   1.940   0.0553 .
`2014-4`    -3.076e-03  1.436e-03  -2.142   0.0347 *
`2014-6`     2.807e-06  1.103e-06   2.544   0.0125 *
`2014-7`     9.384e-04  1.011e-03   0.928   0.3557
`2014-8`     3.096e-06  2.054e-05   0.151   0.8805
`2014-9`    -6.104e-03  5.105e-03  -1.196   0.2347
`2014-10`    1.326e-10  1.068e-10   1.242   0.2171
`2014-11`   -1.058e-03  4.686e-04  -2.258   0.0262 *
`2014-12`   -4.761e-04  1.058e-03  -0.450   0.6536
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.02251282)

    Null deviance: 2.9220  on 106  degrees of freedom
Residual deviance: 2.2063  on  98  degrees of freedom
  (70 observations deleted due to missingness)
AIC: -91.671

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = hub_score ~ ., data = countrywhole2014[, c(9, 13,
    15:21)])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.21722  -0.07056  -0.02723   0.02261   0.82022

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.149e-01  2.677e-01   1.923   0.0573 .
`2014-4`    -3.077e-03  1.476e-03  -2.085   0.0396 *
`2014-6`     2.892e-06  1.134e-06   2.551   0.0123 *
`2014-7`     9.751e-04  1.039e-03   0.939   0.3503
`2014-8`     3.279e-06  2.110e-05   0.155   0.8768
`2014-9`    -6.288e-03  5.245e-03  -1.199   0.2335
`2014-10`    1.355e-10  1.097e-10   1.236   0.2195
`2014-11`   -1.090e-03  4.814e-04  -2.264   0.0258 *
`2014-12`   -4.986e-04  1.087e-03  -0.459   0.6474
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.02376421)

    Null deviance: 3.0821  on 106  degrees of freedom
Residual deviance: 2.3289  on  98  degrees of freedom
  (70 observations deleted due to missingness)
AIC: -85.883

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = strength.2019 ~ ., data = countrywhole2019[, c(2,
    13, 15:17, 19)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-678.75   -89.00   -35.40    14.31  1168.14

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.375e+02  2.688e+02   0.883   0.3784
`2019-6`     4.759e-03  1.154e-03   4.125 6.16e-05 ***
`2019-8`    -1.752e-02  2.882e-02  -0.608   0.5442
`2019-9`    -5.925e+00  5.396e+00  -1.098   0.2739
`2019-10`    8.939e-07  1.198e-07   7.459 6.82e-12 ***
`2019-12`    2.015e+00  1.048e+00   1.924   0.0563 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 56301.73)

    Null deviance: 13604151  on 153  degrees of freedom
Residual deviance:  8332656  on 148  degrees of freedom
  (23 observations deleted due to missingness)
AIC: 2129.4

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = coreness.2019 ~ ., data = countrywhole2019[, c(3,
    13, 15:17, 19)])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-17.0583   -4.8595    0.0474    4.3169   12.8896

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.165e+01  6.930e+00   1.681 0.094898 .
`2019-6`     1.266e-04  2.975e-05   4.255 3.69e-05 ***
`2019-8`     1.752e-04  7.429e-04   0.236 0.813872
`2019-9`    -1.033e-01  1.391e-01  -0.743 0.458782
`2019-10`    1.107e-08  3.090e-09   3.583 0.000461 ***
`2019-12`    1.110e-01  2.701e-02   4.111 6.51e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 37.42533)

    Null deviance: 9054.4  on 153  degrees of freedom
Residual deviance: 5538.9  on 148  degrees of freedom
  (23 observations deleted due to missingness)
AIC: 1002.8

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = page.rank.2019 ~ ., data = countrywhole2019[, c(4,
    13, 15:17, 19)])

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-0.0176699  -0.0022720  -0.0008321   0.0006808   0.0314465

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.556e-03  7.045e-03   0.930   0.3536
`2019-6`     1.256e-07  3.024e-08   4.154 5.50e-05 ***
`2019-8`    -6.161e-07  7.553e-07  -0.816   0.4160
`2019-9`    -1.361e-04  1.414e-04  -0.962   0.3376
`2019-10`    2.332e-11  3.141e-12   7.426 8.21e-12 ***
`2019-12`    4.966e-05  2.746e-05   1.809   0.0726 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 3.868302e-05)

    Null deviance: 0.0092894  on 153  degrees of freedom
Residual deviance: 0.0057251  on 148  degrees of freedom
  (23 observations deleted due to missingness)
AIC: -1119.7

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = betweenness.2019 ~ ., data = countrywhole2019[,
    c(5, 13, 15:17, 19)])

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-867.3  -182.0  -104.7    50.9  1866.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.488e+02  4.312e+02  -0.345  0.73060
`2019-6`     5.858e-03  1.851e-03   3.165  0.00188 **
`2019-8`    -2.169e-03  4.623e-02  -0.047  0.96264
`2019-9`     3.664e+00  8.656e+00   0.423  0.67275
`2019-10`    3.643e-07  1.922e-07   1.895  0.06003 .
`2019-12`    1.877e+00  1.681e+00   1.117  0.26577
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 144895.4)

    Null deviance: 25472611  on 153  degrees of freedom
Residual deviance: 21444518  on 148  degrees of freedom
  (23 observations deleted due to missingness)
AIC: 2275

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = closeness.2019 ~ ., data = countrywhole2019[, c(6,
    13, 15:17, 19)])

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-3.572e-04  -1.082e-04   7.300e-07   1.028e-04   4.611e-04

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.204e-03  1.625e-04   7.408 9.04e-12 ***
`2019-6`    1.331e-09  6.976e-10   1.909  0.05825 .
`2019-8`    1.155e-08  1.742e-08   0.663  0.50843
`2019-9`    2.366e-06  3.263e-06   0.725  0.46945
`2019-10`   1.243e-13  7.245e-14   1.716  0.08826 .
`2019-12`   2.041e-06  6.334e-07   3.223  0.00156 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 2.058269e-08)

    Null deviance: 3.8586e-06  on 153  degrees of freedom
Residual deviance: 3.0462e-06  on 148  degrees of freedom
  (23 observations deleted due to missingness)
AIC: -2280.7

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = eigen_centrality.2019 ~ ., data = countrywhole2019[,
    c(7, 13, 15:17, 19)])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.30154  -0.04223  -0.01862   0.00371   0.86341

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.126e-01  1.514e-01   2.065   0.0407 *
`2019-6`     2.645e-06  6.498e-07   4.070 7.62e-05 ***
`2019-8`    -1.365e-05  1.623e-05  -0.841   0.4015
`2019-9`    -6.805e-03  3.039e-03  -2.239   0.0266 *
`2019-10`    1.990e-10  6.749e-11   2.949   0.0037 **
`2019-12`    8.635e-04  5.900e-04   1.463   0.1455
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.01786054)

    Null deviance: 3.4461  on 153  degrees of freedom
Residual deviance: 2.6434  on 148  degrees of freedom
  (23 observations deleted due to missingness)
AIC: -174.96

Number of Fisher Scoring iterations: 2
```

We are also curious about the relationship between the number of airlines between two nodes and the centrality measurement as well as other socioeconomic variables mentioned above. Same as the node level regression, we ran generalized linear model (GLM) between the number of airlines between the two countries and centrality/socioeconomic characteristics of each country (number of airlines as the y variable).

```
Call:
glm(formula = edges ~ ., data = g14[, c(3, 5, 7:11, 13:19, 21,
    23:27, 29:35)])

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-60.87  -10.15   -0.69    5.40  882.35

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.624e+01  1.717e+01   2.694  0.00711 **
coreness.2014.x      1.742e-01  2.387e-01   0.730  0.46554
betweenness.2014.x   3.199e-03  1.949e-03   1.641  0.10093
closeness.2014.x    -2.864e+04  1.363e+04  -2.102  0.03568 *
eigen_centrality.2014.x 3.762e+04  4.619e+04   0.814  0.41543
authority.x         -1.896e+04  2.341e+04  -0.810  0.41797
hub_score.x         -1.861e+04  2.278e+04  -0.817  0.41409
`2014-6.x`           1.041e-04  5.265e-05   1.978  0.04807 *
`2014-7.x`          -1.207e-01  6.468e-02  -1.867  0.06205 .
`2014-8.x`           1.311e-03  1.016e-03   1.289  0.19732
`2014-9.x`          -4.076e-02  2.021e-01  -0.202  0.84017
`2014-10.x`          8.696e-09  3.609e-09   2.410  0.01603 *
`2014-11.x`         -5.364e-02  2.714e-02  -1.976  0.04822 *
`2014-12.x`          5.278e-02  5.818e-02   0.907  0.36436
coreness.2014.y      1.742e-01  2.387e-01   0.730  0.46554
betweenness.2014.y   3.199e-03  1.949e-03   1.641  0.10093
closeness.2014.y    -2.864e+04  1.363e+04  -2.102  0.03568 *
eigen_centrality.2014.y 3.762e+04  4.619e+04   0.814  0.41543
authority.y         -1.896e+04  2.341e+04  -0.810  0.41797
hub_score.y         -1.861e+04  2.278e+04  -0.817  0.41409
`2014-6.y`           1.041e-04  5.265e-05   1.978  0.04807 *
`2014-7.y`          -1.207e-01  6.468e-02  -1.867  0.06205 .
`2014-8.y`           1.311e-03  1.016e-03   1.289  0.19732
`2014-9.y`          -4.076e-02  2.021e-01  -0.202  0.84017
`2014-10.y`          8.696e-09  3.609e-09   2.410  0.01603 *
`2014-11.y`         -5.364e-02  2.714e-02  -1.976  0.04822 *
`2014-12.y`          5.278e-02  5.818e-02   0.907  0.36436
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1551.954)

    Null deviance: 5617382  on 3063  degrees of freedom
Residual deviance: 4713283  on 3037  degrees of freedom
  (462 observations deleted due to missingness)
AIC: 31236
```

To look at whether there are any correlation between the number of edges between the two countries and the socioeconomic factors of the said two countries, we ran a regression using the country-level data. From the results, we can see that GDP per capita and Population(total) show positive correlations at 0.01 confidence level while closeness and trade-in services (% of GDP) show negative correlations at the same confidence level. Also, individuals using the internet (% of population) shows negative correlations at 0.1 confidence level. The above results suggest that GDP per capita and Population are strongly and positively correlated to centrality at the node level, which intuitively makes sense as more people and higher GDP indicates more demand for air travel from people who can actually afford it. The negative correlation between individuals using the internet (% of population) and number of edges between two countries can be hard to interpret at first, but it makes sense upon a closer look. If we were to sort the results by decreasing percentage, we can see that the top countries are dominated by wealthy small northwestern European countries. These countries typically have a lot of wealth and well-built internet infrastructure. Individuals in these countries typically have really high demands for traveling. However, this effect is mostly offset by population, in which these wealthy small European countries have a huge gap compared to countries like China and India. Despite having less wealth and percentage of population using the internet, the sheer size and population of these countries drive demand for more flight paths. Thus, the internet factor is overpowered by population factors and appears negative for the entire network.

Implications of Our Results & Interesting Possible Extensions

Before we jump into analyzing the results of our analysis, let's first summarize what we have done and look at their implications. We started off by visually inspecting the global network by the airport level, but found that the network was too convoluted for any meaningful results. We then moved down to the city level in an attempt to simplify things, before finally settling in around the country level, where our networks struck the perfect balance between being too complex for pattern finding and too simple for meaningful findings. We then took subsets of our data by each continent and again produced the similar three-level graphs as mentioned above. We noticed that for some continents, the city level produced better networks (e.g. Oceania) while for other continents the country level was still the right level of complexity (e.g. Europe or Asia). From these graphs, we can conclude that regions can be drastically different from each other and that further analysis may be needed on only the continental levels. This is especially true when we take into account socioeconomic factors or simply geography. For example, Europeans might rely less on air travel due to the close proximity of the countries, while Asians might heavily depend on air travel due to the island nature of southeast Asia. Caution should be exercised when conducting blanket analysis on the entire international network. We then moved onto some preliminary exploratory data analysis, including finding the top 5 most popular domestic and international routes, as well as the most popular cities and countries (popularity is measured by the number of routes in both cases). These top rankings give us some initial expectations as to which airports, cities or countries should we be on the lookout for, and serves as a proxy for which nodes might exhibit higher centrality measures. We

extended our analysis of this concept by calculating the centrality measures of all three levels (airport, city, and country).

Through coreness, betweenness and eigenvector centrality, we can see each node's centrality figures, which gives us an idea of the airports closest to the core of the network, the airports who act as important bridges between large separated parts of the network, and the airports with the largest influences respectively. These allow us to find the key airports that are critical for the way goods and people are currently moving around the world. For example, Anchorage International Airport's high betweenness is a fantastic example of an airport that people normally wouldn't expect when it comes to being a critical node in the global air transportation network. Not only is it a major air freight hub by connecting Asia with North America, but it is also an important passenger hub that connects remote communities around Alaska that depend on air travel due to the topographic nature of the state. As a finishing touch, we looked at each node's authority and hub measures. Interestingly, most airports did exhibit high asymmetry between the two measures, with the largest being at 0.05 for the UK at the country level, at 0.02 for New York at the city level, and at 0.02 for Kunming International Airport in China at the airport level. We know that a hub is a node that links many other nodes with very frequent routes and that an authority is a node that is being linked to by many authority nodes. To further our analysis of these measures, we can look at how these nodes transform over time, and whether they always held these important positions, or has their centrality measures shifted over time. While the asymmetry is low for 2014 data, we can look at how this asymmetry evolves over time, and if it grows or shrinks for particular nodes at particular levels.

Zooming back to the network level, we then looked at normalized network-level centrality measures of closeness, betweenness, and degree of the three levels of the global network (airport, city, and country again.) We noticed that the airport level by far had the lowest measures of the three levels, with the country level being typically many times higher. This intuitively makes sense, as airport-to-airport differences are typically minimal, while country-to-country differences can be huge. High deviations in geographic, socioeconomic, political and many other factors could have great influences on where and how many airports and routes a country could have. We then conducted the same analysis again, but with data from 2019. By comparing the two sets of results, we can see that airport-level measures all increased, suggesting that airport types, destinations and uses have all diversified. On the city level, network-level closeness increased while network-level betweenness and degree both stayed relatively similar. This suggested that some cities have become closer to each other, while other cities became farther apart, which hints at the formation or solidification of cliques. On the country level, all three measures slightly reduced, meaning that countries have become more similar, which could be the result of globalization influences. On the topic of cliques, we then looked at how the biggest cliques in the network evolved over time. The largest cliques in both 2014 and 2019 were in Europe, with the airport-level being 20 airports, city-level being 21 cities and country-level being 18 countries in 2014. In 2019, the size of the cliques increased on all three levels, increasing to 24, 25 and 22 for airport, city, and country respectively. With this being set in Europe, this could mean that air travel is becoming more prevalent, and more inclusive. This could be influenced by migratory factors, the most notable of which is the large influx of refugees from the Middle-East in recent years. Lastly, we conducted several regression

analyses using node-level centrality measures of the country network and socioeconomic factors such as GDP, population density and other factors highlighted in part 4. In total, we ran 8 regressions using the directed 2014 data and 6 using the undirected 2019 data. While we are unable to derive hub and authority measures from the 2019 data, we noticed some interesting trends from the other measures. Predominantly, GDP and population size tend to be significant and highly correlated with most centrality measures. Because there is so much analysis that can be done using these centrality measures and most factors under the sun, we purposely chose a broad selection of factors that represent different areas that people typically consider (e.g. economics, population, development, education level, etc.). Extending research on this aspect is seemingly endless, and we suggest people find the most significant factors and select a few sub-factors from that specific category.

While we have conducted ample analysis already, we would like to some possible extensions to other topics. Firstly, we suggest finding more data to distinguish among the edges in our data. By finding out the composition of the types of travelers for each route (e.g. leisure, business, etc.), we can identify each node's importance for different types of travel at every level. Just like how we looked at asymmetry between each node's hub and authority measures, we can also look at potential asymmetry between each node's degree and inbound/outbound number of passengers and freight. This can help us analyze the composition of the aircrafts at each node, as well as the size and type of each node. By then creating simulations where we remove select airports, we can see how the entire network and transportation of people and goods are affected, allowing us to create contingency plans to maximize transport efficiency and minimize congestion. that could become vital in times of natural disaster. Of course, we can also

conduct further regression analysis on factors such as geographical features (latitude, longitude, altitude, timezone, etc.) to see if there is a preference in travel destinations, and help us potentially predict demand of new nodes. We can also look at correlation between the price of the tickets of each route and their centrality measures in order to analyze price trends and if some routes are overpriced/underpriced compared to their peers. Our dataset also includes the airline that operates each route, and some really interesting insights can be drawn from looking at the nationality of the airlines of each edge. This can give us a proxy of the balance of power between the two nodes of each edge, and whether one airport, city or country is more reliant on this edge than the airport, city or country on the other side of the edge. With the regions divided up, we can conduct cross-continent comparison between their centrality measure and see characteristics that are unique to each region and how they are influenced by local factors. Lastly, we also suggest looking at current political anomalies (e.g. Trump's presidency, trade wars/agreements, etc.) and their effects on flight paths over time, which could be really useful in predicting price influxes and potentially creating ticket price hedging and arbitrage opportunities.

As Carl picked up his suitcase and headed to his boarding group line, Carl pondered about the future of the global air travel network. Has it already reached a mature stage, or is it still under development? How will the network evolve over the next decade or two? If space travel one day becomes the next air travel network, will it exhibit similar characteristics? Carl is humbled by how much the world has changed thanks to air travel, and can't help but feel thankful for having taken Social Network Analytics, which has equipped him with the tools to understand such a fascinating yet complicated system.

Appendix

We have included all our analysis results in a Google Drive folder for your convenience.

Please find it here.

Select slides from our presentation:

Work Cited

Open Flights

https://openflights.org/

World Bank Socio-Economic Data

https://data.worldbank.org/

Igraph documentation
https://igraph.org

Hub and Authority model discussion by Easley and Kleinberg
https://www.cs.cornell.edu/home/kleinber/networks-book/networks-book-ch14.pdf

2014 IATA Global Passenger Survey
https://www.iata.org/publications/Documents/2014%20IATA%20Global%20Passenger%20Survey%20Highlights.pdf

More IATA Data
https://www.iata.org/publications/Pages/index.aspx

2019 Survey Sneak Peek
https://www.iata.org/publications/store/Documents/IATA-2019-GPS-Highlights.pdf

U.S. International Air Passenger and Freight Statistics Report
https://www.transportation.gov/policy/aviation-policy/us-international-air-passenger-and-freight-statistics-report