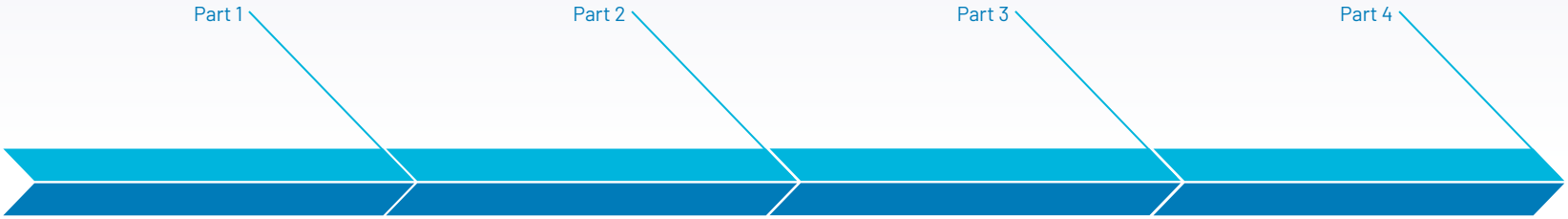
EMORY GOIZUETA BUSINESS SCHOOL | Master of Science in Business Analytics MSBA

Social Network Analytics Final Project

Team Excited

Carl Xi, Jie Zhu, Meng Cheng

Presentation Flow



Project Inspiration

Why are we excited about planes???

Data Exploration

What is exciting about our data???

Analysis Insights

What exciting stuff did we find from our data???

Further Research

What exciting things can we do with our findings???

1

Project Inspiration

Why are we talking about planes???

> 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

~4,000,000,000

Passengers travel over a global network of some 22,000 routes

Problem Statement

If we were to use airports as nodes and flight routes as a proxy for edges, what node, group, or network level effects can we observe?

Problem Statement

In particular, with this dataset, we expect to answer questions such as:

- What are the most popular destinations?
- Is this popularity associated with geographical measures (latitude, longitude, altitude, time zone)?
- Which cities/countries are important in the global network?
- Does our network have any correlation with socioeconomics factors?

2014 Flight Data

- Using global flight paths as proxy connections
- More charters = stronger bond

2

Data Exploration

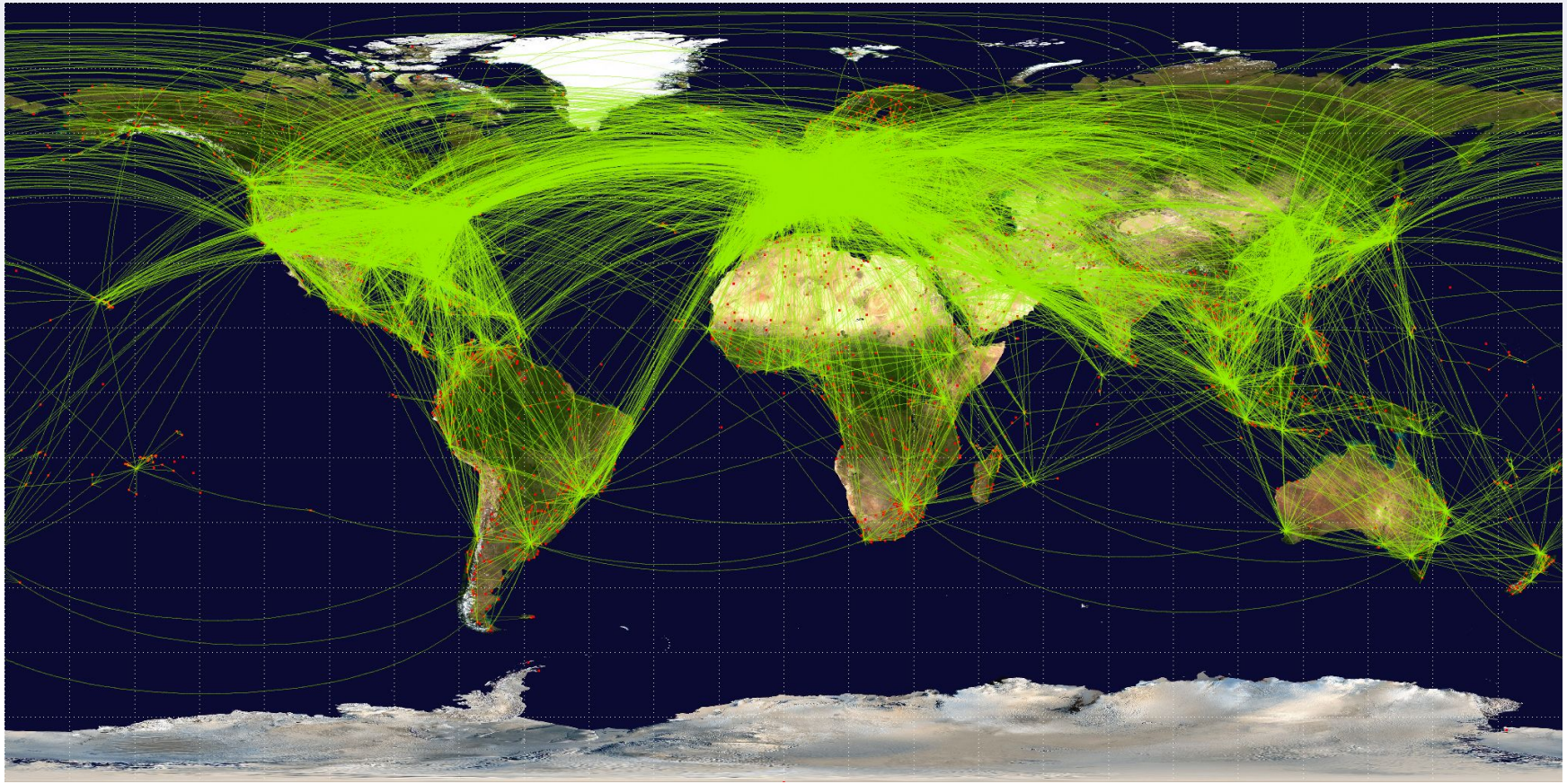
Dataset Introduction &
Exploratory Data Analysis

Dataset Introduction

- Source: OpenFlights.org
- Two useful databases
  - Route database
    - Last Update: June 2014
    - Contains information of routes operated by commercial airlines across the world
  - Airport database
    - Last Update: January 2017
    - Contains information of airports and their city & country of affiliation

Route Database

- **Airline** 2-letter (IATA) or 3-letter (ICAO) code of the airline.
- **Airline ID** Unique OpenFlights identifier for airline.
- **Source airport** 3-letter (IATA) or 4-letter (ICAO) code of the source airport.
- **Source airport ID** Unique OpenFlights identifier for source airport (see Airport).
- **Destination airport** 3-letter (IATA) or 4-letter (ICAO) code of the destination airport.
- **Destination airport ID** Unique OpenFlights identifier for destination airport (see 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



Example: PVG to LAX

American Airlines

Codeshare with other airlines, not the actual operator

```
> head(route[Source_Airport == 'PVG' & Destination_Airport == 'LAX'])
   Airline AirlineID Source_Airport Source_AirportID Destination_Airport Destination_AirportID Codeshare Stops Plane
1:      AA        24            PVG             3406                 LAX                  3484                    0   777
2:      CA       751            PVG             3406                 LAX                  3484          Y         0   788
3:      DL      2009            PVG             3406                 LAX                  3484          Y         0   346
4:      HU      2660            PVG             3406                 LAX                  3484          Y         0   777
5:      MU      1758            PVG             3406                 LAX                  3484                    0   346
6:      UA      5209            PVG             3406                 LAX                  3484                    0   788
```

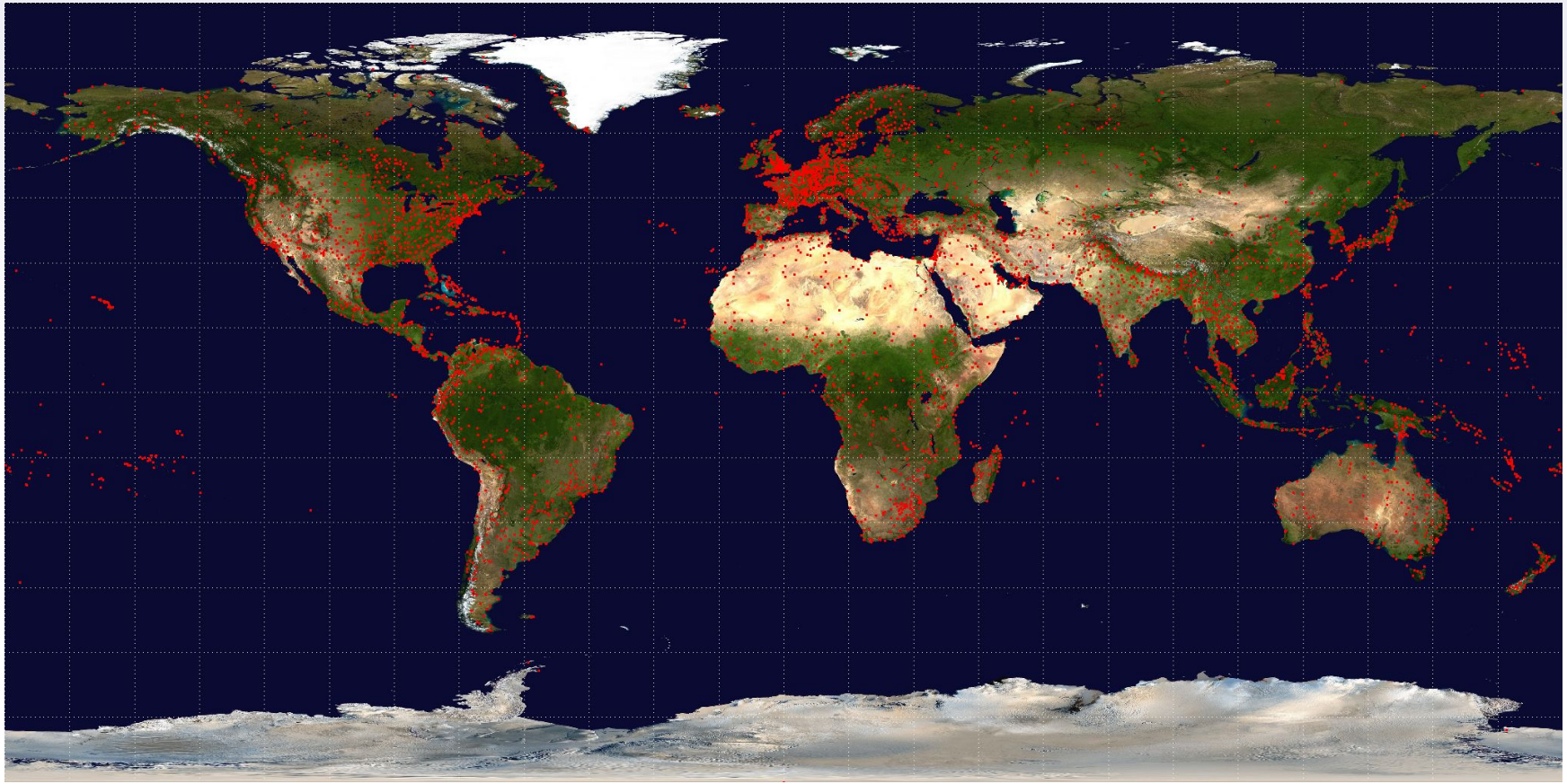
IATA code for Shanghai Pudong International Airport

IATA code for Los Angeles International Airport

346 stands for Airbus 340-600
788 stands for Boeing 787-8

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 airport is located.
- **IATA** 3-letter IATA code. Null if not assigned/unknown.
- **ICAO** 4-letter ICAO code.
- **Latitude** Decimal degrees, usually to six significant digits. Negative is South, positive is North.
- **Longitude** Decimal degrees, usually to six significant digits. Negative is West, positive is East.
- **Altitude** In feet.
- **Timezone** Hours offset from UTC. Fractional hours are expressed as decimals, eg. India is 5.5.



Merged Dataset

- Merge on **Airport ID** for both **Source Airport** and **Destination Airport**
- NAs after merging Route Database and Airport Database
  - 406 NAs in **Source City** and 416 NAs in **Destination City**
  - Some small airports are not listed in the Airport Database
  - Keep routes with NAs for now, because they are still useful in calculating other airports' centrality measures
- Primary Variables
  - Source Airport / City / Country
  - Destination Airport / City / Country
- Other Variables
  - Airline, Codeshare
  - Altitudes, Latitudes, Longitudes, Time Zones

37,595

Unique Routes (53.9% Int.)

3,425

Unique Airports

568

Unique Airlines

Top Domestic Routes

	Country	Routes
1st Place	United States	5,895
2nd Place	China	5,698
3rd Place	Brazil	1,175
4th Place	Russia	942
5th Place	Canada	932

Top International Routes

	Origin	Destination	To	Back
1st Place	United Kingdom	Spain	491	485
2nd Place	German	Spain	332	331
3rd Place	Mexico	United States	247	245
4th Place	Germany	Italy	171	171
5th Place	China	Taiwan	164	164

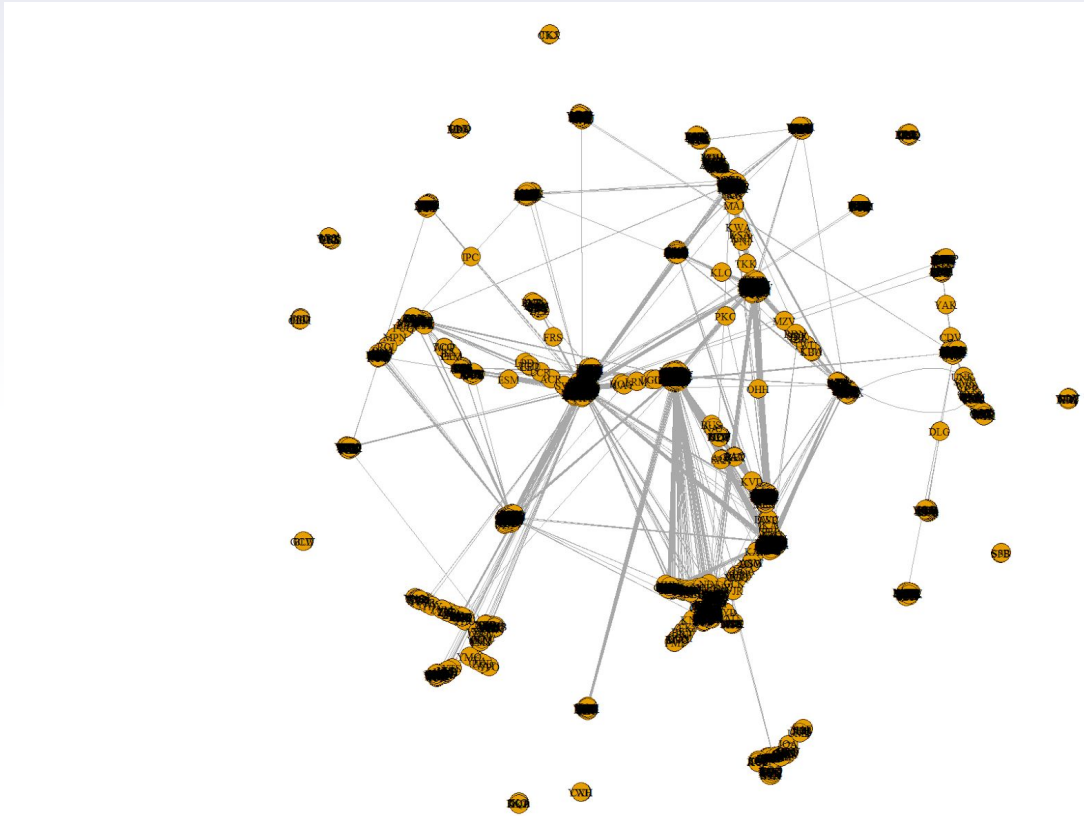
Top City Routes

	Origin	Destination	Routes
1st Place	Shanghai	Taipei	17
2nd Place	Seoul	Tokyo	15
3rd Place	Taipei	Tokyo	15
4th Place	London	Palma de Mallorca	14
5th Place	Arrecife	London	13

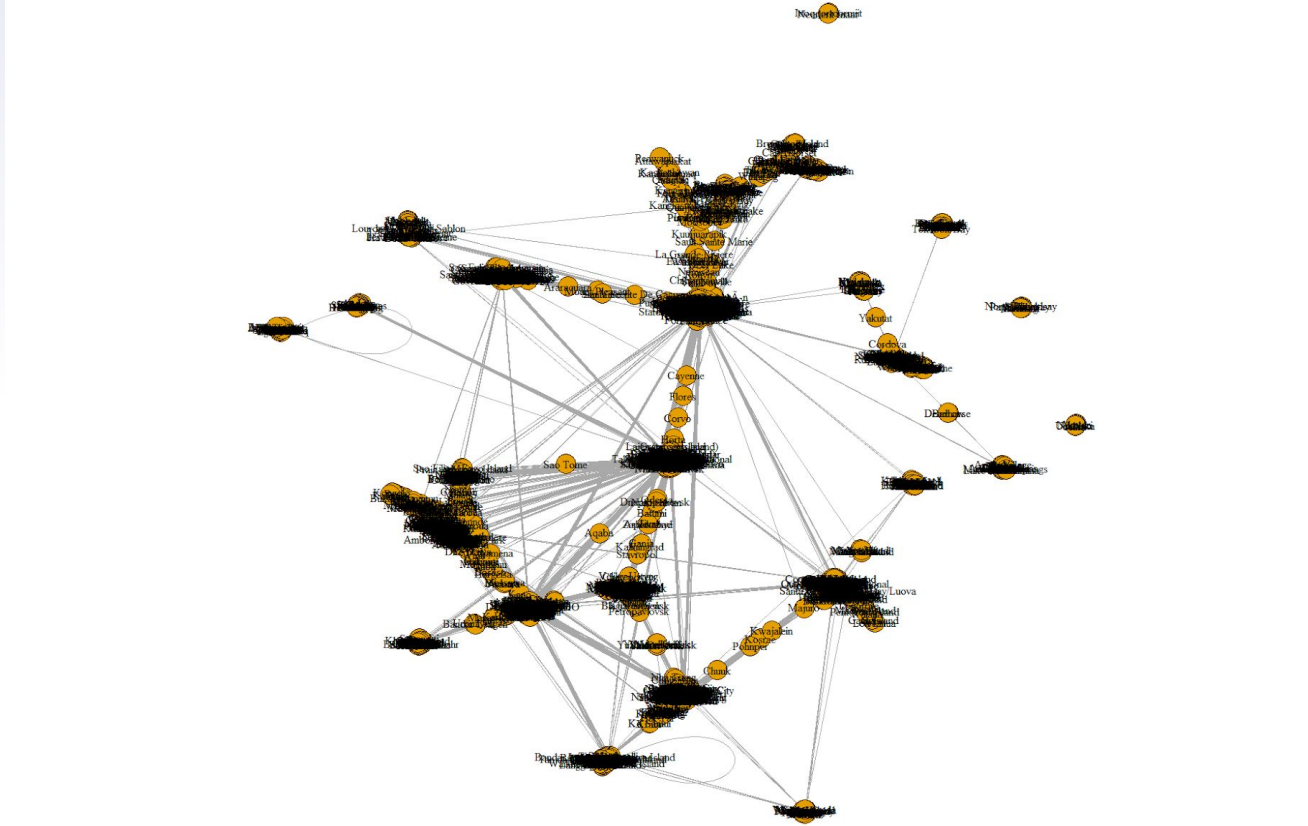
3

Analysis Insights

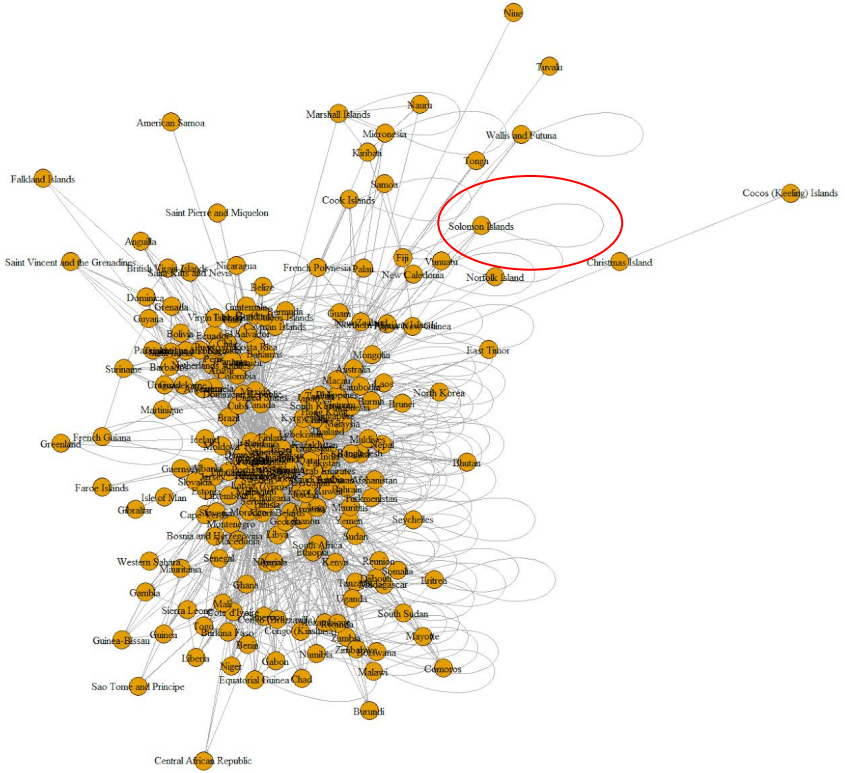
What insights can we drive?



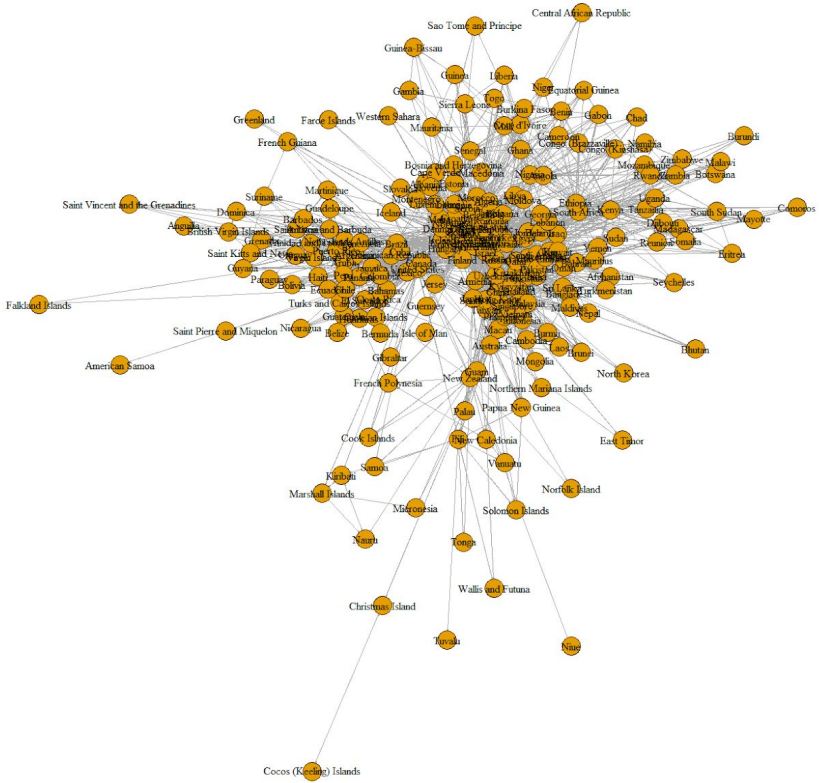
Full Raw Network



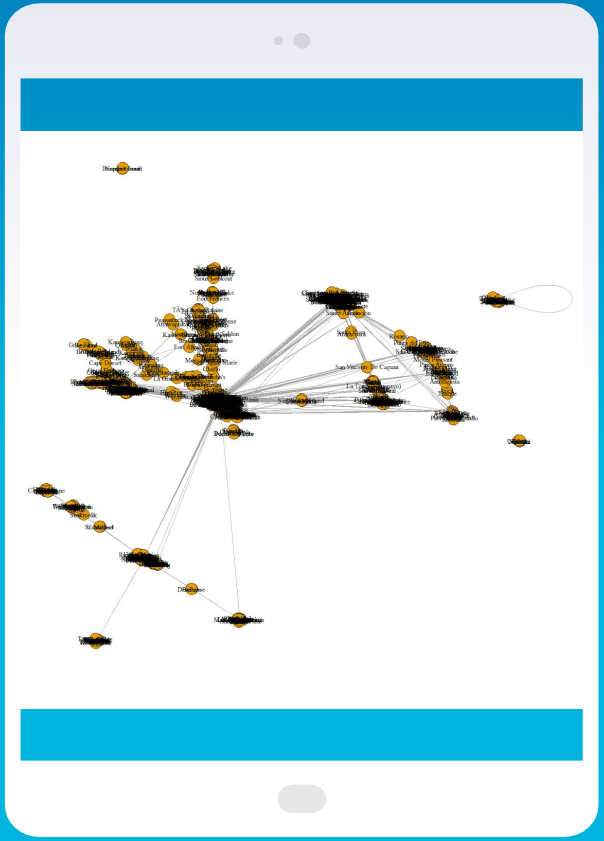
Cities Network



Countries Network

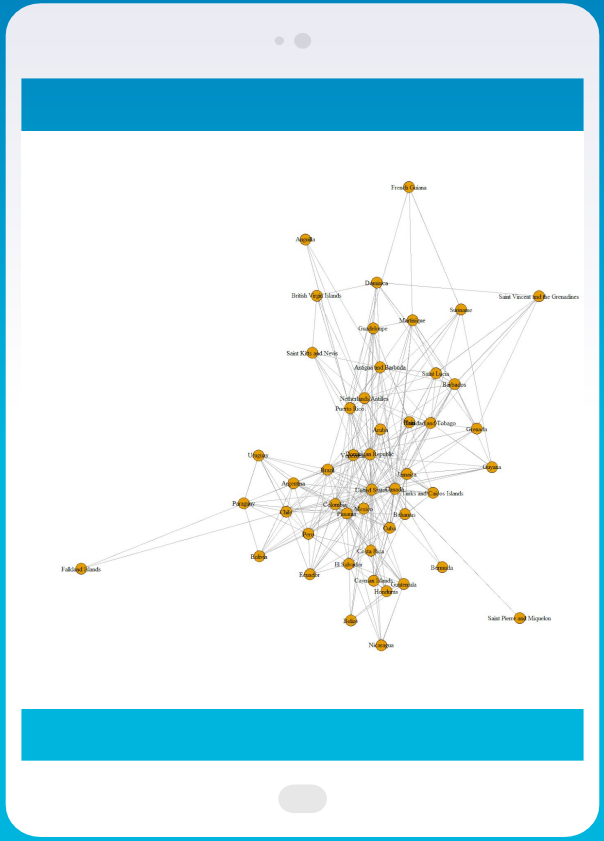


Countries Network*



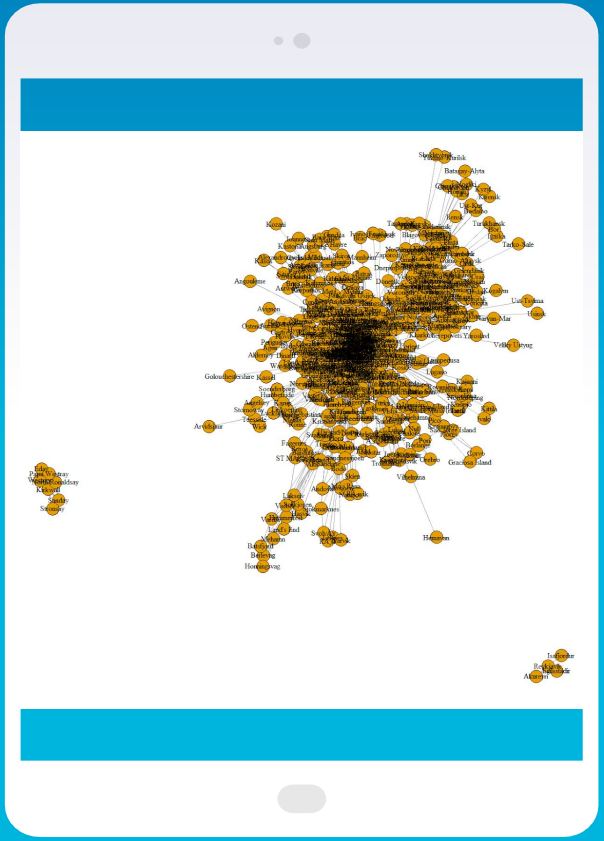
City Level

Country Level



Country Level (Int. Only)

Region: Americas

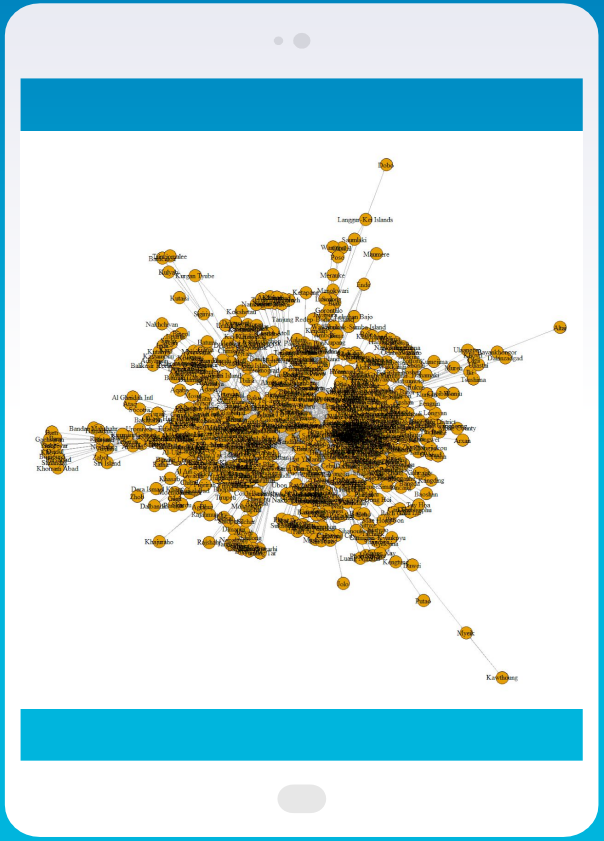


City Level

Country Level

Country Level (Int. Only)

Region: Europe

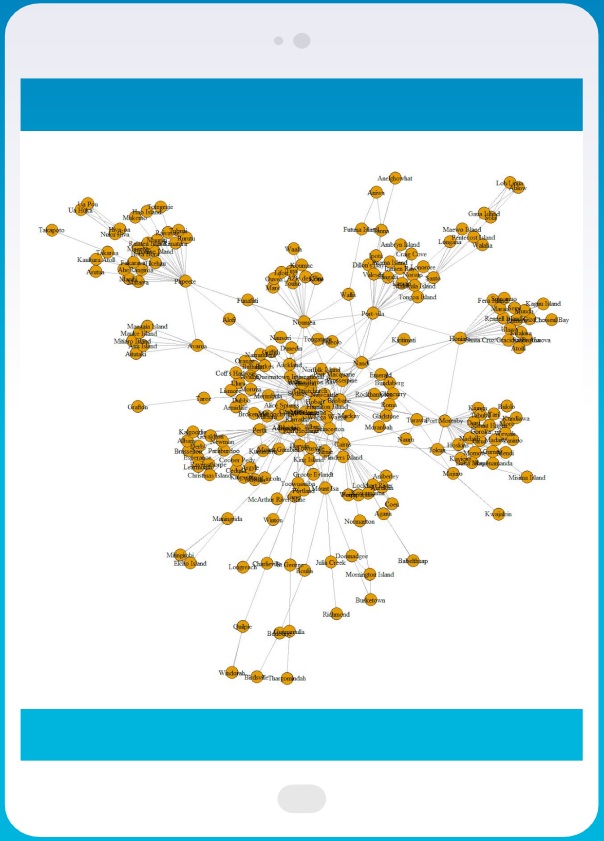


City Level

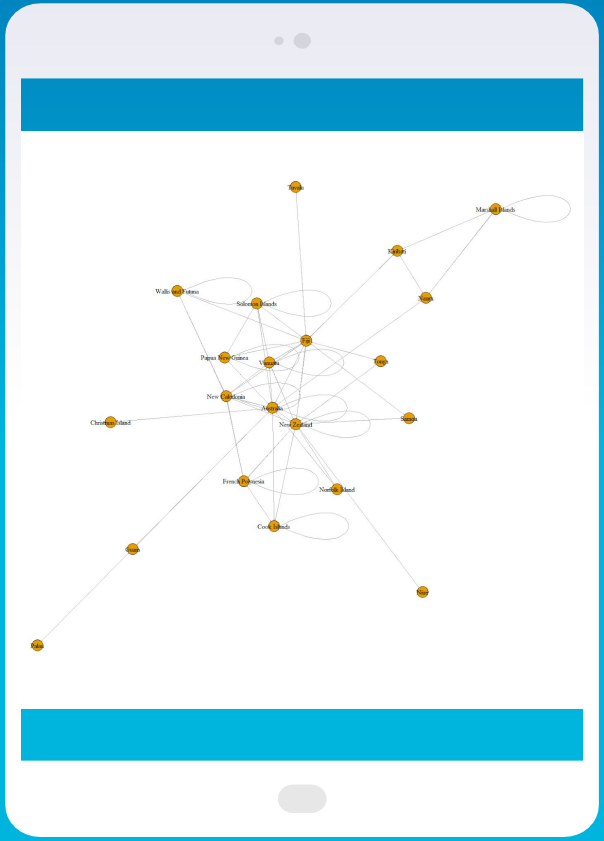
Country Level

Country Level (Int. Only)

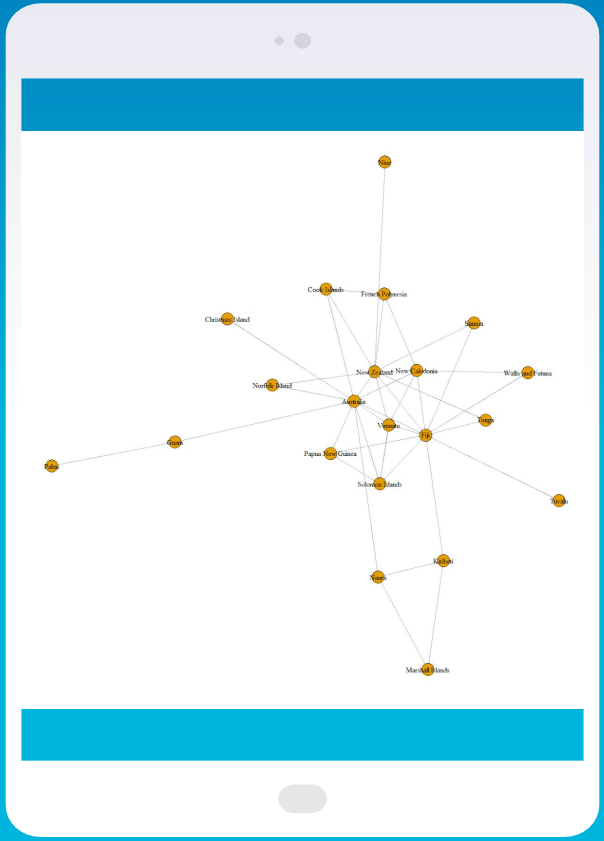
Region: Asia



City Level

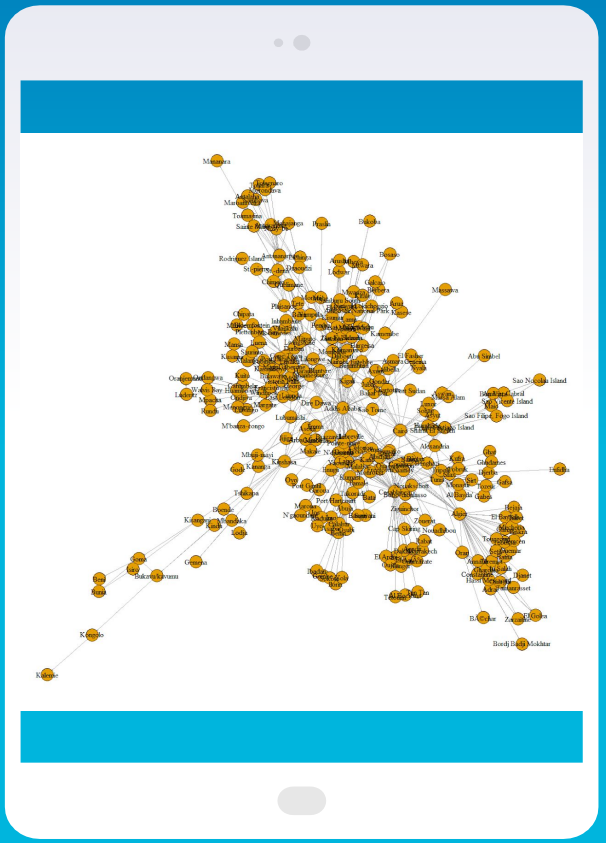


Country Level

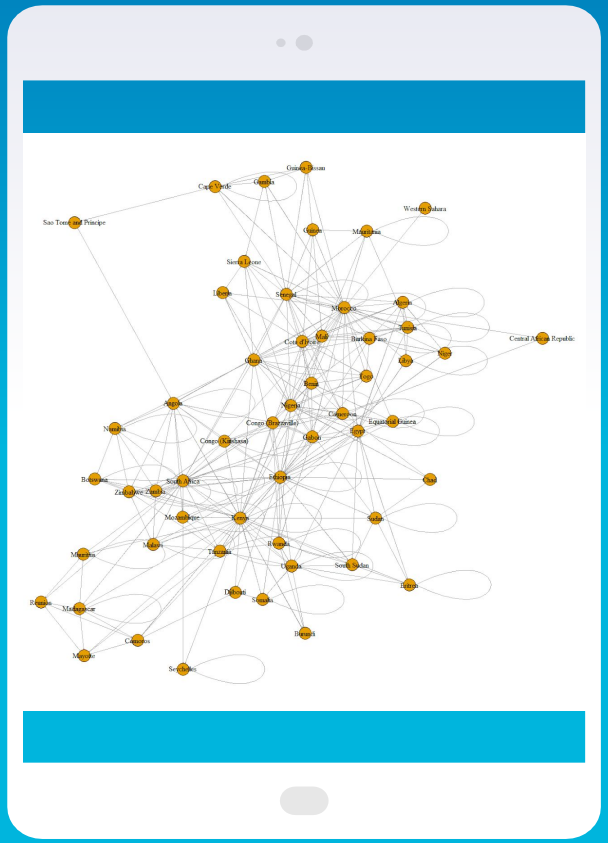


Country Level (Int. Only)

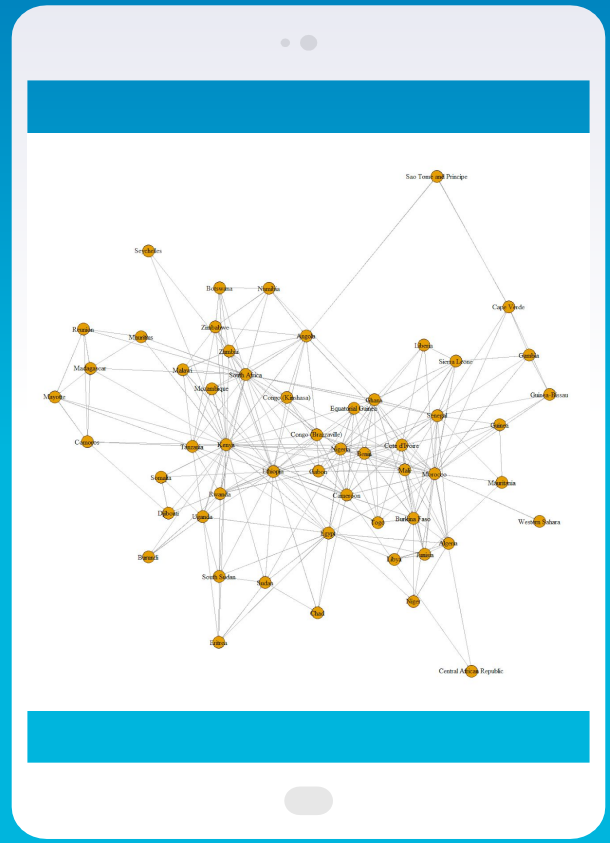
Region: Oceanic



City Level



Country Level



Country Level (Int. Only)

Region: Africa

Early Results

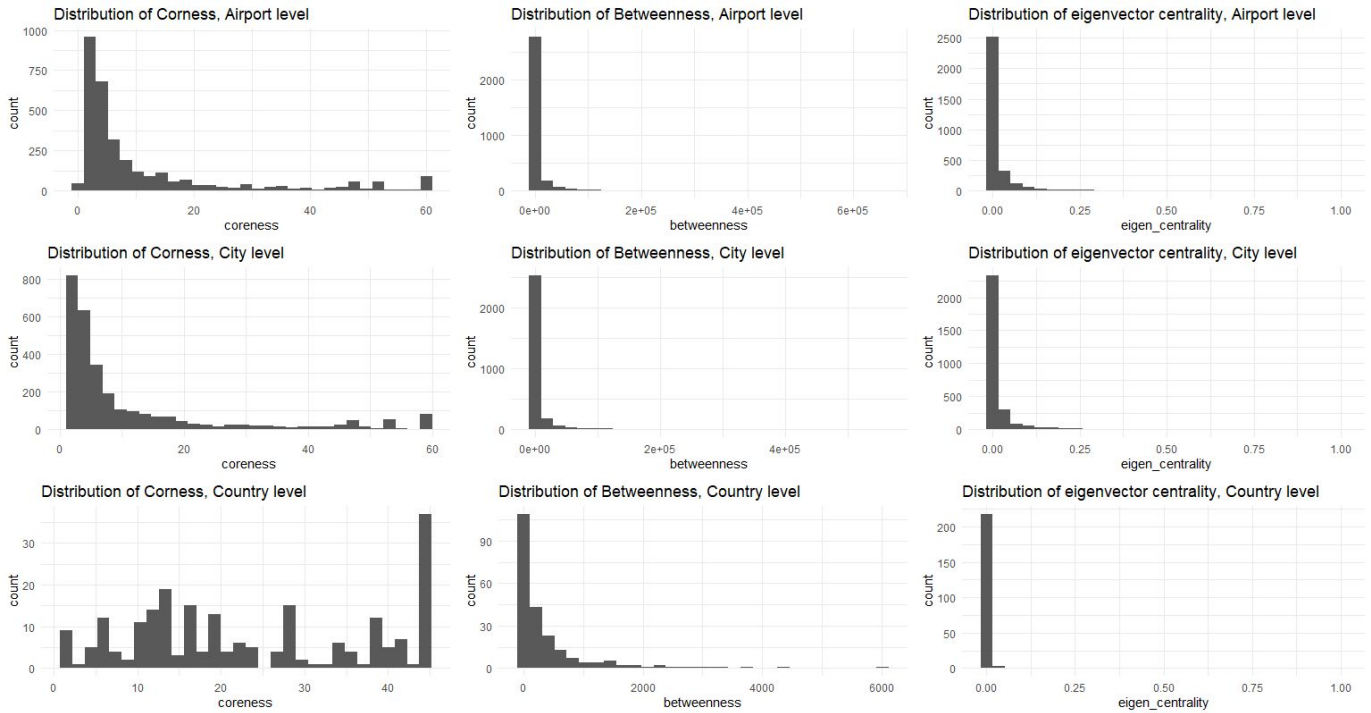
Node Level

- Betweenness
- Coreness
- Eigen Centrality
- Degree
- Closeness
- Page Rank

Group Level

- Cliques

Centrality Figures



Top Airports

	Coreness	Betweenness	Eigen Centrality
1st Place	CDG	ANC	PVG
2nd Place	FRA	AMS	PEK
3rd Place	IST	FRA	HKG
4th Place	LGW	DXB	ICN
5th Place	AMS	NRT	CAN

Top Cities

	Coreness	Betweenness	Eigen Centrality
1st Place	London	Tokyo	London
2nd Place	Moscow	Frankfurt	Shanghai
3rd Place	Paris	Amsterdam	Tokyo
4th Place	Istanbul	Paris	Beijing
5th Place	Dubai	London	Paris

Top Countries

	Coreness	Betweenness	Eigen Centrality
1st Place	United States	France	United States
2nd Place	China	Canada	China
3rd Place	United Kingdom	Qatar	Mexico
4th Place	Spain	Ethiopia	Canada
5th Place	Germany	Netherlands	United Kingdom

Regression Results -Economic Factor

Coreness/Betweenness/Eigenvector Centrality/PageRank ~ GDP Per Capita + Population

	Coreness	Betweenness	Eigenvector Centrality	PageRank
GDPP	+ ***	+ ***	+ *	+ ***
POP	+ ***	+ .	+ **	+ ***

Significant level (P-Value) 0 = '***', 0.001 = '**', 0.01 = '*', 0.05 = '.'

Network Level Centrality

With Normalization

	Airport Level	City Level	Country Level
Closeness	0.0008	0.0013	0.0923
Betweenness	0.0655	0.1054	0.1717
Degree	0.0703	0.1142	0.4247

Clique

Find largest clique with size 20, 21, 18 at Airport, City and Country level.

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18
France	Germany	United Ki	Italy	Turkey	Ireland	Spain	Denmark	Poland	Greece	Belgium	Switzerlar	Norway	Sweden	Malta	Latvia	Czech Rep	Hungary
France	Germany	United Ki	Italy	Turkey	Ireland	Spain	Denmark	Poland	Greece	Belgium	Switzerlar	Norway	Sweden	Finland	Latvia	Czech Rep	Hungary
France	Germany	United Ki	Italy	Turkey	Ireland	Spain	Denmark	Poland	Greece	Belgium	Switzerlar	Norway	Sweden	Finland	Croatia	Czech Rep	Hungary
France	Germany	United Ki	Italy	Turkey	Ireland	Spain	Denmark	Poland	Greece	Netherlan	Switzerlar	Norway	Sweden	Malta	Latvia	Czech Rep	Hungary
France	Germany	United Ki	Italy	Turkey	Ireland	Spain	Denmark	Poland	Greece	Netherlan	Switzerlar	Norway	Sweden	Finland	Latvia	Czech Rep	Hungary
France	Germany	United Ki	Italy	Turkey	Ireland	Spain	Denmark	Poland	Greece	Netherlan	Switzerlar	Norway	Sweden	Finland	Croatia	Czech Rep	Hungary
France	Germany	United Ki	Italy	Turkey	Russia	Switzerlar	Belgium	Spain	Denmark	Poland	Greece	Israel	Finland	Sweden	Latvia	Czech Rep	Hungary
France	Germany	United Ki	Italy	Turkey	Russia	Switzerlar	Belgium	Spain	Denmark	Poland	Greece	Norway	Sweden	Malta	Latvia	Czech Rep	Hungary
France	Germany	United Ki	Italy	Turkey	Russia	Switzerlar	Belgium	Spain	Denmark	Poland	Greece	Norway	Sweden	Finland	Latvia	Czech Rep	Hungary
France	Germany	United Ki	Italy	Turkey	Russia	Switzerlar	Belgium	Spain	Denmark	Poland	Greece	Norway	Sweden	Finland	Croatia	Czech Rep	Hungary
France	Germany	United Ki	Italy	Turkey	Russia	Switzerlar	Netherlan	Spain	Latvia	Poland	Greece	Israel	Finland	Sweden	Denmark	Czech Rep	Hungary
France	Germany	United Ki	Italy	Turkey	Russia	Switzerlar	Netherlan	Spain	Latvia	Poland	Greece	Norway	Sweden	Denmark	Malta	Czech Rep	Hungary
France	Germany	United Ki	Italy	Turkey	Russia	Switzerlar	Netherlan	Spain	Latvia	Poland	Greece	Norway	Sweden	Denmark	Finland	Czech Rep	Hungary
France	Germany	United Ki	Italy	Turkey	Russia	Switzerlar	Netherlan	Spain	Finland	Poland	Norway	Greece	Denmark	Sweden	Croatia	Czech Rep	Hungary

4

Further Research

Where can we go from here?

Taking a Step Back

Node Level

- Degree
- Betweenness
- Closeness
- Coreness
- Page Rank
- Eigen Centrality

Group Level

- Cliques
- Transfer Effect

  (Directional Preferences)

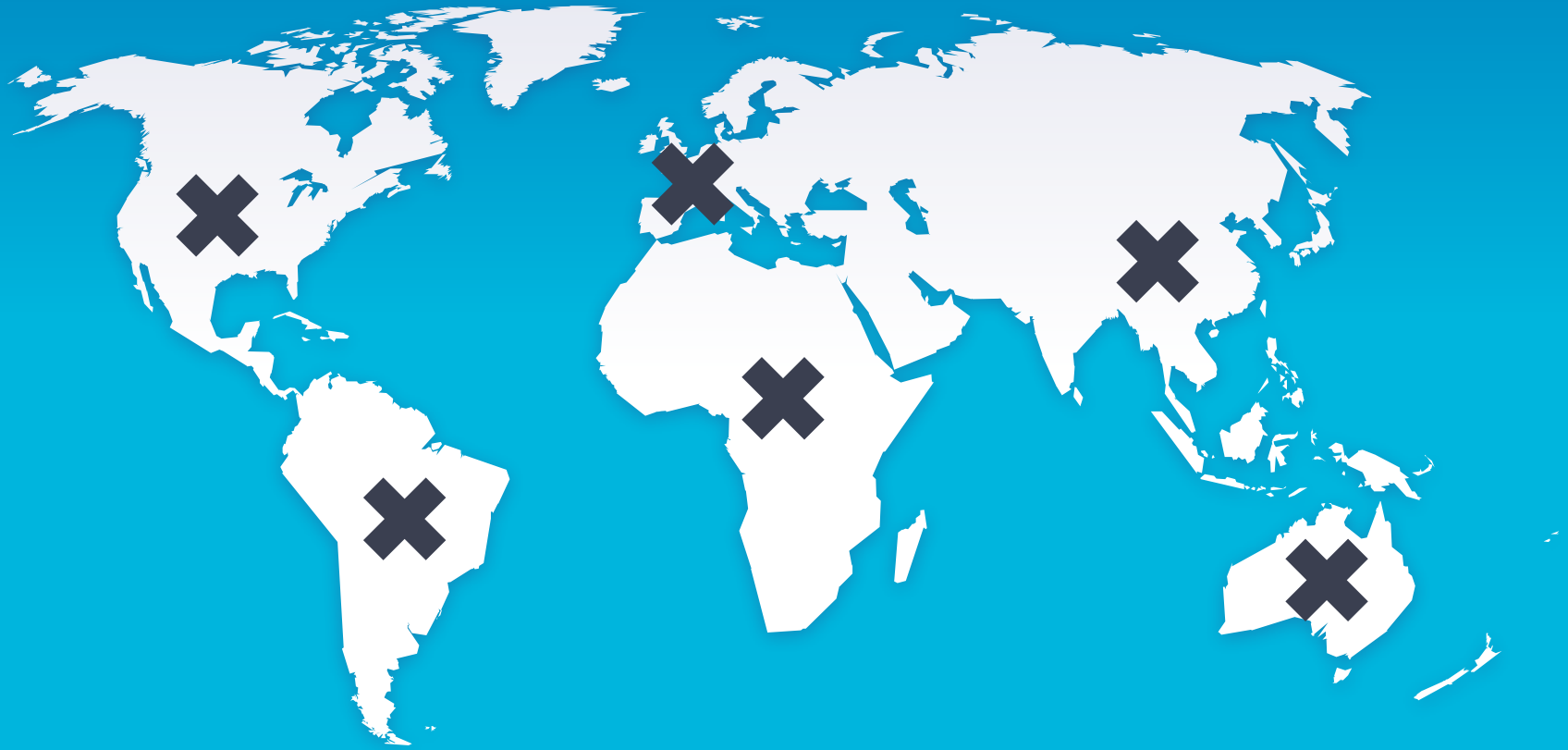
Network Level

- Formation & Stability of ties in evolution of network
- Time series data

Betweenness = importance?

- ANC (Anchorage International Airport) has the highest betweenness of all airports in the world
- Topographic map - many regions are only reachable via air

Regional Cliques?



Flight Path Nationality

- Looking at international flight paths can help us evaluate reliance
  - E.g. majority of New York - London flights are British airlines.
    - This could hint at British reliance on US-UK relationships.

Next Steps

Cross-Region Comparisons

- We can categorize our data by continent, region or country, re-conduct our entire analysis process and cross-compare the results

Further Regression

- We can run the various network measures (e.g. centrality figures) against all sorts of socioeconomic factors (e.g. we used GDP as an example in our analysis)

Time Series Analysis

- A lot interesting insights can be derived from looking at how the network changed over time

Using our results as a Springboard

Our EDA & Initial Analysis

Our analysis lays a solid foundation for further experimentations

Further Hypothesis Testing on all 3 levels

We recommend using our results and asking questions around the summary statistics & social network measures we derived

Data Prediction & Prescription

Combined, we can further test correlation & causation relationships between our network and socioeconomic factors to potentially predict & prescribe solutions for current global problems

THANKS!

Any Questions?

Please also let us know if you have any comments, concerns, or recommendations!

Appendix

- EDA

```
#unique value involved
uniqueAirline <- unique(data$Airline) #568
uniqueRoute <- unique(data[,c(1,2)]) #37595
uniqueAirportD <- unique(data$Destination_Airport) #3418
uniqueAirportS <- unique(data$Source_Airport) #3409
uniqueAirport <- unique(append(uniqueAirportD,uniqueAirportS)) #3425
uniqueCountry <- unique(append(data$Country_S,data$Country_D)) #226
uniqueCity <- unique(append(data$City_S,data$City_D)) #3142

#codeshare involved
nshare <- sum(data$Codeshare=="Y") #14597

#data remove codeshare route
datac <- data[which(data$Codeshare!="Y"),]
nrow(datac) #53066

#number of international routes
InternationRoute <- datac[which(datac$Country_S!=datac$Country_D),] #28604
nrow(InternationRoute)/nrow(datac) #53.9%
```

Appendix

- Regression Results

```
Call:
lm(formula = coreness ~ GDPP + pop, data = regc)

Residuals:
     Min       1Q   Median       3Q      Max
-24.3827  -9.0983  -0.3367   8.4860  20.8134

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.162e+01  1.129e+00  19.151  < 2e-16 ***
GDPP        3.217e-04  4.194e-05   7.671 1.77e-12 ***
pop         2.699e-08  7.968e-09   3.388 0.000893 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.09 on 155 degrees of freedom
Multiple R-squared:  0.3054,    Adjusted R-squared:  0.2964
F-statistic: 34.07 on 2 and 155 DF,  p-value: 5.438e-13
```

```
Call:
lm(formula = betweenness ~ GDPP + pop, data = regc)

Residuals:
    Min      1Q  Median      3Q     Max
-2368.8  -293.6  -201.4    95.9  4926.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.005e+02  8.105e+01   2.474   0.0144 *
GDPP        1.850e-02  3.011e-03   6.142 6.59e-09 ***
pop         1.016e-06  5.721e-07   1.777   0.0776 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 796.2 on 155 degrees of freedom
Multiple R-squared:  0.2051,    Adjusted R-squared:  0.1948
F-statistic: 19.99 on 2 and 155 DF,  p-value: 1.885e-08
```

Appendix

- Regression Results(cont)

```
Call:
lm(formula = eigen_centrality ~ GDPP + pop, data = regc)

Residuals:
     Min       1Q   Median       3Q      Max
-0.19680 -0.00783  0.00042  0.00406  0.92067

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.095e-03  7.838e-03  -0.905  0.36677
GDPP         6.588e-07  2.912e-07   2.262  0.02509 *
pop          1.576e-10  5.533e-11   2.849  0.00498 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.077 on 155 degrees of freedom
Multiple R-squared:  0.07566,   Adjusted R-squared:  0.06373
F-statistic: 6.343 on 2 and 155 DF,  p-value: 0.002249
```

```
Call:
lm(formula = page.rank ~ GDPP + pop, data = regc)

Residuals:
      Min        1Q    Median        3Q       Max
-0.024811 -0.001861 -0.000757  0.000369  0.077790

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.565e-03  8.299e-04   1.885   0.0612 .
GDPP        1.559e-07  3.084e-08   5.055 1.20e-06 ***
pop         3.211e-11  5.858e-12   5.482 1.67e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.008153 on 155 degrees of freedom
Multiple R-squared:  0.2557,    Adjusted R-squared:  0.2461
F-statistic: 26.63 on 2 and 155 DF,  p-value: 1.147e-10
```

Appendix

- Network Measure Calculations

```
# network-level measure
#Airport Level
centr_clo(AirportG,normalized=TRUE)$centralization #0.0008350746
centr_betw(AirportG,normalized=TRUE)$centralization #0.06548108
centr_degree(AirportG,normalized=TRUE)$centralization #0.07034328


#City Level
centr_clo(CityG,normalized=TRUE)$centralization #0.001333726
centr_betw(CityG,normalized=TRUE)$centralization #0.1054135
centr_degree(CityG,normalized=TRUE)$centralization #0.1142897

#Country Level
centr_clo(CountryG,normalized=TRUE)$centralization #0.09233869
centr_betw(CountryG,normalized=TRUE)$centralization #0.1717404
centr_degree(CountryG,normalized=TRUE)$centralization #0.4247119
```

Appendix

- Clique

```
#cliques #ingnore directionality
clique_num(AirportG) #20
clique_num(CityG) #21
clique_num(CountryG) #18
```