# CS224W Final Project: Predicting Airline Alliance Composition

Rafael Setra and Austin Hou

November 17, 2016

## 1 Introduction

Airline networks are a fairly recent phenomenon - though the first one was created almost a century ago, large global alliances did not appear until the last couple decades. Alliances can be forged for monetary, political, or other reasons. In this project we build a network model of the three largest airline alliances, and attempt to construct a metric by which to identify ideal non-alliance candidates for induction into an existing alliance, as well as to evaluate existing airlines in a given alliance.

When forging airline alliances, many factors are considered on both the side of an alliance and the side of the individual airline. Advantages on the alliance side include integration of international and regional networks, retention of loyal customers through integrated loyalty programs, unified branding and marketing, and optimization of route scheduling. Advantages on the airline side include access to worldwide marketing, maintenance, and support resources; access to international markets and increased exposure to high value customers; and expansion of codeshare and route accessibility. Here we take the perspective of a large alliance and focus on regional and international route integration and member airline ”fit”.

We want to identify ”good fit” airlines for alliances by constructing a metric through which we can analyze and quantify the routes a specific airline serves, as well as the benefit that these routes would provide to an existing worldwide network. We will apply this algorithm to airlines both within and outside an alliance. This information can benefit a prospective airline that desires to enter negotiations with an alliance, and an alliance by quantifying the efficiency and integration of it’s network.

## 2 Related Work

### 2.1 Graph Evolution [1]

Graph evolution over time is an important aspect to consider when developing graph generation models. One particular type of growth that we will focus on is the average distance between nodes. The authors showed that networks exhibit a decreasing diameter as it grows over time.

For a graph $G$, they investigate the full diameter $d$, the maximum length of a path between any two connected nodes, and the effective diameter $D = g(G)$. This value is such that the fraction of connected nodes of distance $D$ or less is 0.9. This value $D$ can be extended to non integer values by a linear interpolation of the integer values about it. This value can be computed relatively quickly [2]. The authors also described graph generation models which have the decreasing diameter property.

There was a clear trend observed in the diameter measured for the networks analyzed in this paper. However, it could be difficult to generalize this trend to other networks, in particular air transportation since the majority of networks covered were related to human interaction. It would be nice to see data for other networks composed of businesses or institutional relationships, as these types of networks could be significantly different from collaboration networks.

## 2.2 Air Transportation Network [3]

In this paper the authors characterize the global airline network as a small world network. They analyze the roles of different cities and routes using betweenness and degree centrality, and find that the distribution of betweenness in regards to an airline network roughly obeys the functional form $P(> b) \approx b^{-v}$ where $P(> b)$ is the probability of a node's betweenness being above $b$ and $v = 0.9 \pm 0.1$ is the power law exponent.

The authors divide the cities of the world into communities, using modularity. They then analyze the connectedness of communities within their own context, and conclude that the most central cities and the most connected cities are not correlated. They suggest that using this method of community analysis, it may be possible to identify poorly connected communities and suggest new routes to better connect them.

The airline network analysis brings up an interesting metric by which to characterize betweenness. We can clearly see a pattern in the betweenness of airline networks, and describing the characteristics of the network is done well. However, the authors are not trying to predict whether certain conditions should hold for geopolitical structures such as alliances, and we are unsure if this analysis will directly hold to structures aside from organic communities determined by modularity. We propose applying the betweenness power law to alliance commuinities. We will use the ideas presented here in the context of prediction - given a change in the network, determine if the betweenness distribution still follows a power-law.

## 2.3 Link Prediction [4]

In this paper, the authors explore and evaluate a number of methods to predict links in a social network. These methods are divided into three categories - neighbor analysis, path analysis, and more complex higher level approaches. The authors find that because a network is influenced by outside factors, the raw performance that they observe is low, and define a control method consisting of random edge predictions. Comparing the performance of these methods, it was found that a number of techniques performed well, including the Katz, common neighbor, and Adamic/Adar measures. The latter two were surprising given their simplicity. When applied to the small-world problem, the Katz and Adamic/Adar measures performed similarly well, relative to the other measures.

Since we can model airline networks as a small-world problem [3], we can attempt to apply the most successful metrics that were used, such as Katz and Adamic/Adar, and analyze each proposed route individually. We can then combine these ratings by taking their mean across all routes in a prospective airline, and see how they affect the fit of an airline.

# 3 Methods

## 3.1 Data and Network

The data we will be using is a list of routes located at [5]. This is a list of 67663 routes between 3209 airports and 531 airlines since January 2012. For each route, we have the following data: airline, aircraft, source airport, destination airport. We will shorten this dataset to only include the airline and airport information. From this dataset we will create our networks.

We will have three sets of networks, one for each major alliance (Star Alliance, Oneworld, and SkyTeam). For each alliance, we will filter the entire dataset to only include airlines in said alliance (27 airlines in Star Alliance, 16 in Oneworld, and 20 in SkyTeam). Then we will add the routes (edges) that each airline serves between airports (nodes) onto the alliance network. These edges will be unweighted and undirected. Our analysis will then be performed on these three alliance graphs.

## 3.2 Airline Metric

The first step in determining whether an airline is a good fit to an alliance is to establish a metric. The metric that we will consider takes as input the alliance and airline graphs and is of the form of a cost function where positivity signifies a good fit. We were not initially sure which statistics are useful in determine a good airline fit, and so our first step is to determine the importance of the following metrics for the input alliance graph G and airline graph A (G G has removed A):

1. $E(A)$, the number of edges in the airline. More routes may signify an increase in complexity and cost, and also a larger market.
2. $\Delta V = V(G \cup A) - V(G)$, the change in nodes by adding an airline to the alliance. A larger number of airports means the alliance is more global.
3. $\Delta g = g(G \cup A) - g(G)$, the change in effective diameter. We would expect the effective diameter to decrease as in [1] for a more accessible network.
4. $\Delta R^2 = R(G \cup A) - R(G)$, where $R(G)$ is the R-squared value of the line of best fir to the log-log of the between distribution function. According to [3], this distribution should follow a power-law and be linear in the log-log plot.
5. $\Delta LS = LS(G \cup A) - LS(G)$, change in the longest shortest path. We would expect this value to decrease so that overall travel is "short".

6. $Q = Q(y)$, where for each node i, $y_i = 1$ if $i \in A$ and $y_i = -1$ else. This is the modularity. We expect the modularity to increase if the airline has a community structure - a desirable trait.

7. $\Delta Cl = Cl(G \cup A) - Cl(G)$, the change in the clustering coefficient. We would expect a high clustering coefficient so that airports are not isolated.

8. $\Delta CC = CC(G \cup A) - CC(A)$, the change in the average closeness centrality. We would expect the airports to have a higher closeness with an additional airline.

Considering all these metrics, we will consider a cost function of the form:

$$c(G, A) = \sum_{i=1}^{5} c_i X_i.$$

Here, $X_i$ is one of the 8 feature values above, and $c_i$ are constantss. We would like to consider only the five most siginificant metrics for each alliance, and see how they differ. The constants will need to be adjusted depending on which values we want toconsider, which is not an easy decision. The algorithm we decided on was the following:

**Algorithm 1** Feature Selection

```
$A = \{A_1, ..., A_N\}$ a set of airline graphs
$G =$ an alliance network
for $i = 1$ to $N$ do
    Remove $A_i$ from $G$
$F = N \times 8$ zero matrix
for $i = 1$ to $N$ do
    for $j = 1$ to 8 do
        $F_{ij} =$ value of the $jth$ metric from the above list, with current $G$ and $A_i$
    $G = G \cup A_i$
for $j = 1$ to 8 do
    Normalize the $jth$ column of $F$ by its column standard deviation $\sigma_j$.
for $i = 1$ to $N$ do
    for $j = 1$ to 8 do
        $F_{ij} = sign(F_{ij})$
$\hat{F} = \{F_{i_1}, ..., F_{i_5}\}$, $F_{i_i}$ are the columns with the largest absolute column sum

Solve in CVX:
    Let $c = (c_1, c_2, c_3, c_4, c_5)$
    Solve: $Fc > 0$
    Subject to: $\sum_{i=1}^{5} |c_i| = 1$
```

We used CVX in the last step because it ends up being a convex problem since our cost function is affine in the five features.; we need a choice of $c_1, c_2, c_3, c_4, c_5$ such that

$$c(G \cup_{i=1}^{i-1} A_i, A_i) > 0, i = 1, 2, 3, 4, 5.$$

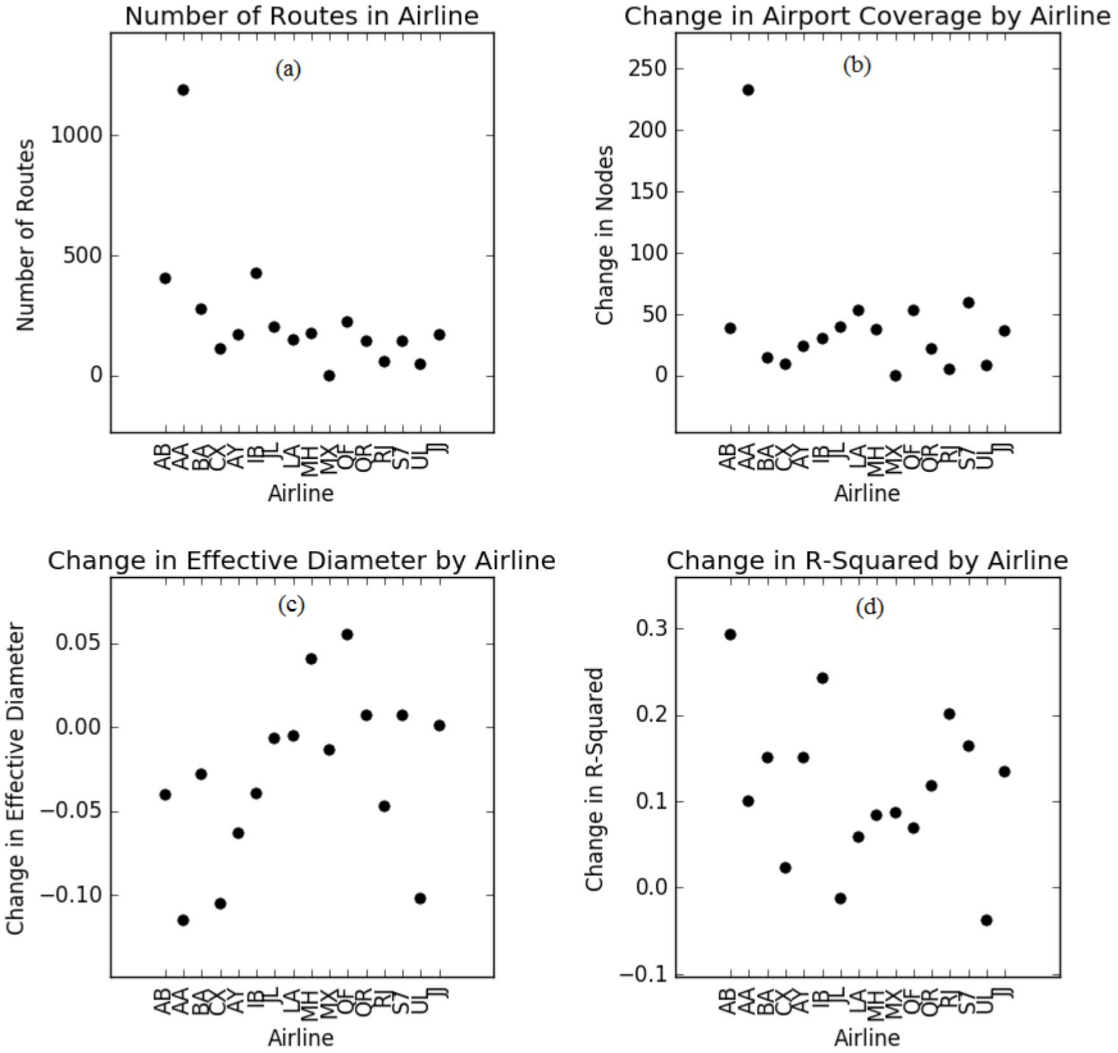
This is solved via entering the conditions into CVX [6]. In practice we ended up with an infinite set of solutions, so we arbitrarily chose the constants over this solution set with the constraint $c_1 + c_2 + c_3 + c_4 + c_5 = 1$ for simplicity. We will run the algorithm above with middle third of $\lfloor \frac{N}{3} \rfloor$ airlines from each alliance network (in order of addition), to result in three vectors of $c$ values. This will let us see what each alliance "values" in selecting an airline. In other words, the $A_i$ will be the middle third airlines; in the case of Star Alliance this is Asiana, LOT Polish, Adria, Croatia, TAP, Swiss, South African, Air China and Turkish Airlines. We will then test these values with the last third airlines in that alliance.

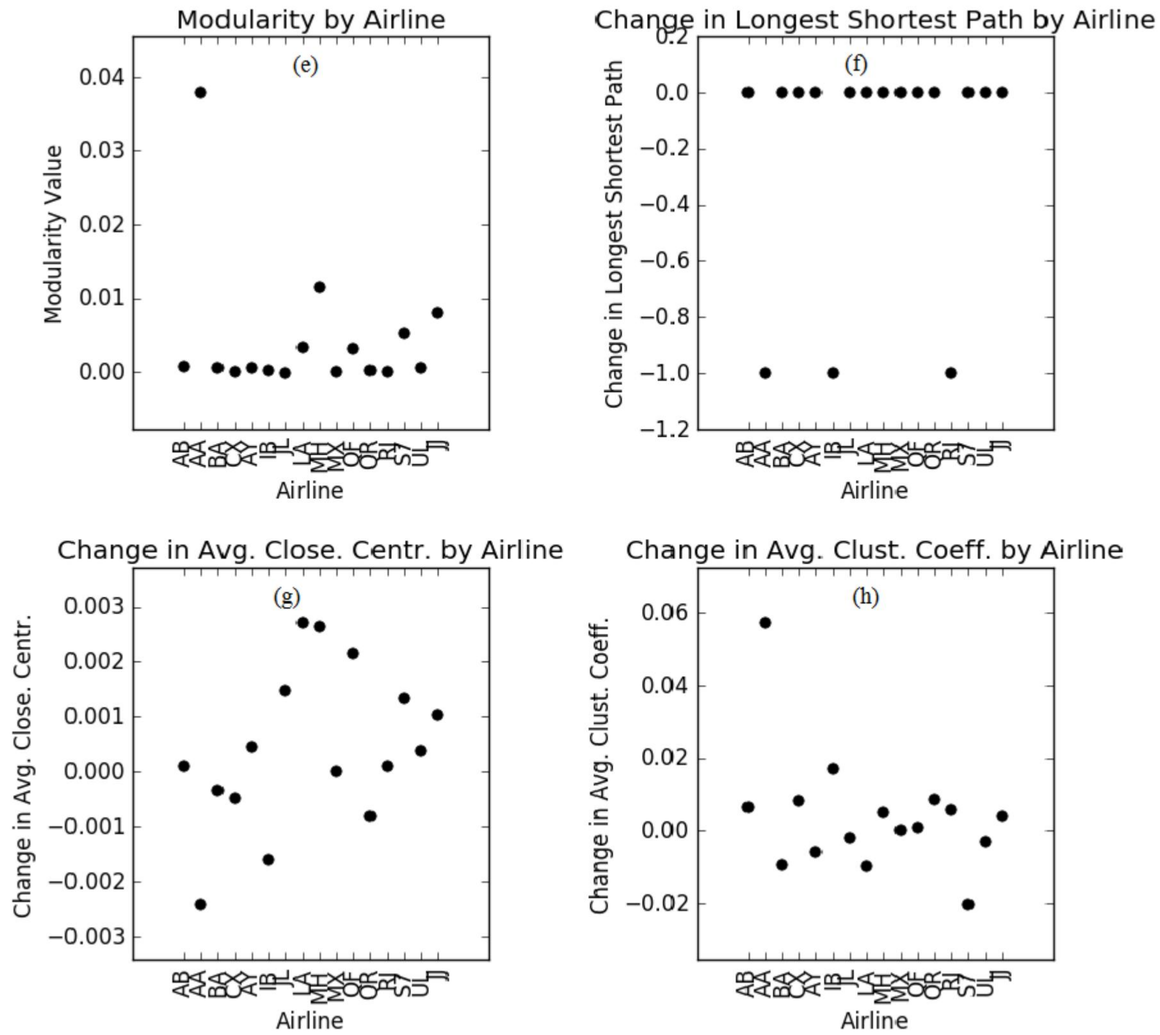
# 4 Results

Now to predict whether an airline should belong to an alliance, we simply removed it from the alliance and found its cost function. If the cost was positive it is a good fit, else it is not.

## 4.1 OneWorld: A Closer Look

Below are figures of the statistics for the Oneworld airlines; note that only a third will be used.





From these statistics, we found that our five most significant features were: $E(A)$, $\Delta V$, $\Delta R^2$, $Q$, and $\Delta LS$. The first two and modularity make sense because those are always positive, and so we would of course pick it. The coefficients of the five significant features are $(c_1, c_2, c_3, c_4, c_5) = (.19, .30, .13, .22, -.16)$ respectively.

From these statistics and our cost function, we found that Sri Lankan(UL) was the only inadequate fit airline onto the Oneworld Alliance, from the last five added. This is because the R-squared value of this airline did not fit in well compared to the other airlines (see Figure (d) ). In addition, it had relatively low airport coverage and routes. This makes sense in our considerations. Also, this was the most recent addition to OneWorld. At this point, the airline may have already been large enough, and the merge may have been for mostly political reasons.

Finally, we attempted to suggest new airlines onto the existing Oneworld alliance by finding airlines with high cost function values, when compared to the full OneWorld network. From the other two alliances, the three highest values are from United Airlines (value of 1.53), Delta Airlines (1.33), and Shenzhen Airlines (1.21).

### 4.2 Star Alliance Summary

From these statistics, we found that our five most significant features were: $E(A)$, $\Delta V$, $Q$, $\Delta LS$, and $\Delta Cl$. The coefficients of the five significant features are $(c_1, c_2, c_3, c_4, c_5) = (.19, .28, .21, -.17, .15)$ respectively. Through our cost function, we found that Avianca and EgyptAir are not good fits into the alliance. This is possibly because the change of closeness centralities due to these two are negative, probably due to their relatively large size - these two offer routes to over 100 airports, while the airlines tested were below 50. The coefficient attached to closeness centrality is positive, so this results in a drop in our cost function.

Finally, we attempted to suggest new airlines onto the existing Star Alliance network by finding airlines with high cost function values, when compared to the full Star Alliance network. From the other two alliances, the three highest values are from Qatar (value of 1.48), Delta Airlines (1.30), and Cathay Pacific (1.12).

### 4.3 SkyTeam Summary

From these statistics, we found that our five most significant features were: $E(A)$, $\Delta V$, $Q$, $\Delta LS$, and $\Delta CC$. The coefficients of the five significant features are $(c_1, c_2, c_3, c_4, c_5) = (.21, .30, .23, -.16, .10)$ respectively. From this alliance, we found that only Middle East Airline was not a good fit into the alliance. This is a relatively small airline with roughly 30 destinations, the average tested in this alliance is about twice that. This makes the values from the edges or change in nodes very small. Additionally, the change in closeness centrality was negative and this brought the cost down below 0.

Finally, we attempted to suggest new airlines onto the existing SkyTeam network by finding airlines with high cost function values, when compared to the full SkyTeam network. From the other two alliances, the three highest values are from American Airlines (value of 1.84), Lufthansa (1.43) and Turkish Airlines (1.37).

## 5 Conclusions

By our results, significant metrics in common across all three alliances were the size of the prospective airline, the number of additional airports served, modularity, and change in longest shortest path. The first three were relatively expected - strong airlines that open new markets and have a good community structure should be desirable to alliances. Longest shortest path was an interesting metric: though it was zero for most of the OneWorld Airlines, a significant proportion of SkyTeam and Star Alliance airlines had nonzero values, and Star Alliance airlines in particular showed a much higher variation here.

Also, each alliance had one metric that was valued lower than the remaining ones. For Oneworld, Star Alliance, and SkyTeam, this was $\Delta R^2$, $\Delta Cl$, and $\Delta CC$, respectively. In each case, these produced the smallest magnitude coefficients, suggesting that they were less significant metrics. This changed for each alliance, showing that each alliance valued different aspects of an airline. However, the three features that are similar in the three alliances (edges, nodes, and modularity) all have very similar weights, showing that they all value these three things about the same, and so the alliances do have some similarity. This could be due to the algorithm we used - they were borderline decisions that could have been swung by a couple data points. In the end, there is not a lot of data to model airline alliances, and some of these inconsistencies could have been moderated by the presence of larger data sets.

Additionally, when we attempted to suggest additions of airlines onto existing alliances, the results showed that our metric favored very large airlines. This makes sense from the $c_i$ values which are fairly large when considering nodes and edges. This also makes sense historically, because the airlines we suggested were initial members in one of the three alliances. In other words, these were considered high valued airlines when the alliances were initially formed, adn this still holds true today.

Lastly, each alliance had fewer than half of the tested airlines fail. This could mean that the metric we have identified is a good measure of what the alliance considered a "good" fit. It would be interesting to define our own metric not based on airlines, but perhaps public opinion and see if the airlines match this new metric.

## 6 Group Contribution

The work was divided evenly between the two members.

## References

[1] J. Leskovec, J. Kleinberg, and C. Faloutsos *Graph Evolution: Densification and Shrinking Diameters*. ACM Transactions on Knowledge Discovery from Data, 1(1), 2007.

[2] C. Palmer, P. Gibbons, and C. Faloutsos *ANF: A Fast and Scalable Tool for Data Mining in Massive Graphs*. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 81-90, 2002.

[3] R. Guimera, S. Mossa, A. Turtschi, and L. A. N. Amaral *The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles* PNAS, 102(22), 2005.

[4] D. Liben-Nowell and J. Kleinberg *The Link-Prediction Problem for Social Networks*. Journal of the American Society for Information Science and Technology, 58(7):10191031, 2007.

[5] Route DataBase
https://raw.githubusercontent.com/jpatokal/openflights/master/data/routes.dat

[6] Matlab Software for Disciplined Convex Programming http://cvxr.com/cvx/