Applying Link Prediction to Global Alliance Networks

CS 224W, Fall 2016

Gus Liu, Bryan Cheong, Zhengyuan Ma

Overview

The alliances and relations between nations determine the course of history. These relations lend themselves quite well to network-analytical studies, as global alliances can be pretty easily modeled as a network linking countries which are allied to each other. Through network analysis, insights into the behavior of nations can be studied just as we may study the activity of individuals over a social network. Unfortunately, these studies have heretofore been few and far in-between, including those answering the basic question: given pairs of countries, can we predict whether or not an alliance will form between them? In this paper, we explore this question, as well as others, such as, "Which features cause countries to be more likely to form alliances with each other?" and "How does the network structure of the graph affect alliance formation?" We implement a prediction system that achieves fairly accurate results which indicate that features such as country contiguity, past disputes, religion, and trade are effective predictors of alliance formation. Moreover, we analyze our results to draw inferences about which countries our system produces false positives for, indicating a shadow alliance, a pair of countries between which it would be reasonable for future alliances to form.

Introduction

Formal alliances and treaties are the legal and political foundations of the present international order. Previous literature such as "Networks of military alliances, wars, and international trade" by Jackson and Nei [1] posit the theoretical game-theoretic stability of the the relationship between countries in terms global warfare and alliances. Jackson and Nei developed a networks model for warfare in which connected allies in a network unite to conquer less powerful nations. They show that under their model, no system of international alliances is stable and there is always a nation with an incentive to change an alliance or attach another nation, which is similar to historical developments in the 20th century.

The pattern of alliance formations is therefore an important and mutable feature of the study of international relations. If we are able to meaningfully predict the formation of alliances between countries, we can inform the discourse on what the future network of global alliances may look like. Thus, we would like to be able to predict possible changes in world alliance networks, given characteristics such as a country's majority religion and geographical location in the world. To this end, the main aim of this paper is to test if we are able to predict and construct the network of international alliances based on certain other properties about the relationship between each country.

The features with which we used to predict the relationship between countries are 1) similarity in religion, 2) scale of former militarised conflicts, 3) physical adjacency, 4) the flow of trade between the two countries. These features are relatively objective and quantifiable using publicly available data, and are descriptions not only of the similarity and differences between any two countries, but are descriptive of the way they had actively pursued their relationship through material means, whether negatively through conflict, or positively through trade.

The first part of the project involves individual prediction of edges, i.e. given two countries and their historical, geographical and social relationship with each other, are they likely to be in an alliance? We explore two approaches to answering this question. The first involves a naive Support Vector Machine (SVM) that treats each edge as a discrete data point, while the second invokes network dynamics.

The second approach is inspired in part by Guha, Kumar, Raghavan, and Tomkins' paper on "Propagation of Trust and Distrust" [2] which developed a model of how trust and distrust propagate through a network of individuals. To summarise the relevant part of their approach briefly, Guha et al.'s model took into account that triads in networks of people often had certain patterns, one of which was that trust is transitive, with the result that many of these triads were monochromatic (i.e. if individual A trusts individual B, and individual B trusts individual C, individual A trusts individual C). We hypothesized that the alliance relationships between countries would have similar behavioural patterns as trust relationships between individuals, in the sense that certain patterns of propagation and consistency within a network of alliances functions like a network of trust. Analogously, an ally of an ally is likely to be an ally.

We thus tested a second predictor: whether we could accurately countries to form alliances if their current allies are predicted to form alliances with each other. We ran two tests: First, for every pair of countries, we iterated over the second country's allies and predicted whether an alliance would form between each ally of the second country and the first country, running our predictor from the first approach as a subroutine. If our predictor predicted that over half of the allies of our first country would form an alliance with the second country, we predicted that the two countries would form an alliance in this model. In the second test, we ran the same test except that we ran our predictor with each ally of the first country paired with each ally of the second country. If over half of the ally pairs were predicted to form an alliance, we predicted that our two countries would form an alliance. We found we could not strongly predict alliances between countries given predicted alliances between their allies.

Model and Data Sets

The network graph we are examining is the relationship between countries, where each country is a node, and each edge contains information about the relationship between any two countries.

Our datasets are currently drawn from the Correlates of War (CoW) project by the University of Michigan, whose stated aim is to "to facilitate the collection, dissemination, and use of accurate and reliable quantitative data in international relations."[3] We use five datasets, and since the CoW project collects data for many years, we used only data collected in the year 1989, to give a snapshot of the world's alliances in this point in history, a point in time before the end of the Cold War and the collapse of the Soviet Union and other significant changes to the geopolitical network landscape.

The *Formal Alliance dataset* forms the response data for our training and test sets. The *Direct Contiguity dataset* provides us with information about a country's geographical location relative to other countries. With this dataset, we label the geographical closeness between countries, with each edge containing information about whether the countries are contiguous by land or closely situated by sea. The *World Religions dataset* contained information about the prevalence of each major religion in a country in question. We used it to see if two countries had the same majority religion, and labelled the edges between them likewise. The *Bilateral Trade dataset* was used to obtain trade volume between countries. The resulting trade volumes were used to label our graph, normalised to describe the amount of trade between any two countries. Finally, the *Militarized Interstate Dispute dataset* contained data to label edges indicating whether there were historical military conflicts between any two countries in the last hundred years, and also included information about the highest instance of CoW's 10-level 'intensity index' for any war that had occurred between the two countries. This final labelling by the CoW is the most subjective feature of our datasets, since it is dependant on CoW's manual weighting of the conflict based on fatalities and the involvement of troops. Otherwise, all our feature data are for most part objective, and based on quantifiable and publicly verifiable data.

Methodology/Algorithms

Our initial approach was predicting an edge formation using an SVM classifier, which, given an input matrix of size *n x m* – where *n* is the number of pairs of nodes in our training set and *m* is the number of features – trains an SVM model using a Gaussian or Radial Basis Function (RBF) kernel, a measure of the squared distance in the feature space between points.

We used the scikit-learn package in python to accomplish this task. We split our data into 70% training data, 30% test data. Then, we constructed our input matrix to our classifier by iterating over pairs of country nodes in our alliance network and constructing feature vectors row by row, with data extracted from the Direct Contiguity, Militarized Interstate Dispute, Trade, and Religion datasets. Our input *y*-labels were 1 if an edge existed in the alliance network and 0 otherwise. After training our model, we ran the predict function on our test data, formatted in the same way, and compared our predictions with the true value of edge existence in the network.

We evaluated our prediction with two main approaches. First, we computed the accuracy of our predictions at 91%. After examining our predictions, we noticed many more predictions of non-edges than edges, due to the imbalance of much higher likelihood of no alliance between countries than otherwise. We computed a confusion matrix of our predictions, and confirmed our hypothesis, having obtained a reasonably high true positive rate – though alongside a relatively high false negative rate.

We decided to adapt our initial approach by using a penalized classifier to bias our model to weight higher our minority class, edge formation. After doing so, our confusion matrix showed instead a high false positive rate, with an reduced false negative rate as well. This showed that we were predicting an edge formation much more frequently, presenting a trade-off between false negatives, predicting no edge where one exists in our true y-labels, and false positives, predicting an edge where none exists. The penalized and unpenalized models can thus be used according to prediction need, whether it is more important to correctly predict an edge forming with a higher probability of being wrong or whether it is more important to minimize incorrectness and potentially miss out on guessing edge formations.

Finally, we decided to try a second approach to introduce network structure into our prediction. Our intuition, recalling Guha, Kumar, et. al., was that if a country c_1's allies, which are its neighbors in the network, are likely to form an alliance with a country c_2, then c_1 is likely to form an alliance with c_2. Thus, after training our model in the same way as the initial approach, we also adapted our predict method by classifying in two different ways:

1. Run predictor on all pairs of nodes between c_1's neighbors and c_2, predict alliance if ≥ 0.50 of pairs had alliances predicted.
2. Run predictor on all pairs of nodes between c_1's neighbors and c_2's neighbors, predict alliance if ≥ 0.50 of pairs had alliances predicted.

Results

The accuracy of our resulting predictions may be summarised using the following confusion matrices. The first row of the matrices contains the number of edges without alliances, the second row contains the number of edges with alliances. The first column describes the number of edges that each algorithm predicted predicted a lack of alliances on the edges in the test set, and the second column describes the number of edges that each algorithm predicted a presence of alliances.

Therefore, for each matrix C below, element $C_{0,0}$ represents the number of true negatives, element $C_{1,0}$ represents the number of false negatives, element $C_{0,1}$ represents the number of false positives, and element $C_{1,1}$ represents the number of true positives in the prediction.

Non-Penalized SVM over edges

5335 (88.81%)	72 (1.20%)
465 (7.74%)	135 (2.25%)

This approach had a relatively high rate of false negatives. This might have been the case because the percentage of formal alliances between countries is actually quite small compared to the percentage of edges between countries without formal alliances. The accuracy of this prediction was 91.1%.

Penalised SVM over edges

4387 (73.03%)	1020 (16.98%)
155 (2.58%)	445 (7.41%)

This approach had a much lower rate of false negatives, but with a correspondingly higher proportion of false positives. However, a high rate of so-called 'false positives' may in fact be desirable if our ultimate aim of investigation is to predict which countries may eventually enter into formal alliances in the future given their historical, geographical and social relationship with each other. A 'false positive' is also an expression of how these countries are naturally affinitive even if they are not currently entered into a formal alliance.

Networks Approach 1
(between c_1's neighbors and c_2)

3900 (64.92%)	1507 (25.09%)
110 (1.83%)	490 (8.16%)

Networks Approach 2
(between c_1's neighbors and c_2's neighbors)

2832 (47.14%)	2575 (42.87%)
168 (2.80%)	432 (7.19%)

These two approaches produced much more false positives than the penalized SVM over edges, leading us to believe that it is more difficult to predict harder to predict whether alliances will form using features of a country's allies. We have shown is that using our relatively limited feature space, we are nevertheless able to predict to a significant and high degree of probability (~91%) whether two countries would be in a formal alliance. By using our second, networks-based approach, we achieve a somewhat lesser accuracy.

The main challenge with our approaches was to choose our feature set and prediction method that can both achieve high accuracy and balance tradeoffs between false positives and false negatives. Because obtaining dyadic data was difficult, we used indicator variables to represent similarities in data, such as religious similarity. More specifically, we represented having the same dominant religion as 1 and any different dominant religion as 0 in our matrix.

Analysis of Results

The aim of our project is ultimately to predict future changes in the network of global alliances. To this end, from our 'false positive' results we are able to obtain an edge list of pairs of countries that, according to our model, are likely to enter into formal alliances, but have yet to doso. Using this edge list, we are able to construct a network of "shadow alliances": alliances that are presumably likely and feasible, but have not materialised in the real world. This was done using the SNAP library. This shadow alliance network contains some interesting features and properties. For the purposes of this analysis, we use the results of our second approach (Penalized SVM on edges) which balances the accuracy of minimising the number of false negatives (predicting the absence of an alliance when there is one present) while having a relatively larger number of false positives.

In particular, we note that most countries in the 'false positive' graph have fewer than ten degrees (i.e. fewer than ten predicted alliances that are not extant in reality). Yet, there are extreme outliers that have up to 78 degrees (i.e. 78 alliances that were predicted, but not in effect). The degree distribution of the false positive graph is shown in Fig. 1, showing this extremely skewed distribution, where only a small number of outliers have more than 25 degrees of false positive edges, while most countries have relatively few falsely predicted. This suggests that our current model is robust for most countries, since most of their alliances are accounted for and predicted.

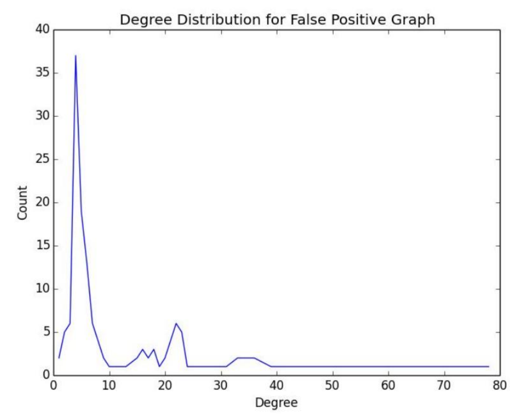

Fig. 1

This is in contrast to the distribution of real-world alliances, which have a higher variance and wider distribution of degrees, as can be seen in Fig. 2. This suggests that the distribution of false positives that we have obtained is fundamentally different from the distribution of extant real-world alliances that our methods otherwise predicted.

From these two observations, we can conclude that our model is mostly robust in predicting the alliances between countries (since the false positive graph has a degree distribution skewed toward nodes with smaller degrees) and is not just the result of randomly drawing from the same distribution as the graph of real-world alliances. From this, we have reason to believe that the nodes with high degrees in our graph of false positives are not countries for which our model has especially high inaccuracy, but are countries that have the highest potential likelihood to form more alliances in the future.

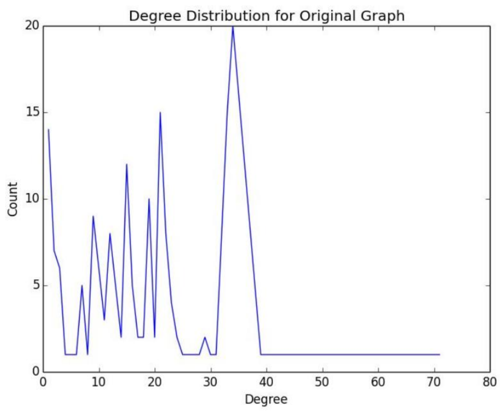

Fig. 2

The Philippines is the country with the highest number of predicted alliances that do not exist, at 78 degrees in the false positive graph. Qualitatively, this makes sense in the geopolitical context, since the Philippines has an export-oriented economy, is a majority-Christian country and has close former colonial ties with the United States, making it a candidate for being in formal alliance with countries in Europe and North America. We suggest that due to historical and political circumstances, such as the fact that the Philippines is a Pacific Rim country far from North America and Europe in the North Atlantic, these alliances have not materialised. Based on our model however, we suggest that the Philippines is a very likely candidate for forming a substantially more connections in the global alliance network in the future.

So far in our analysis, we have been treating countries as though they were passive agents in a connected graph. However, countries are active agents, and can undertake action to affect their influence within the network of global alliances. In particular, a country's connections in the network are a function of its global position and influence. We can see how skewed the distribution of influence between countries are by examining the betweenness centrality of these countries within the global network.

Fig. 3 shows the betweenness centrality distribution of countries in the real-world extant alliance network. Only four countries in the world have a betweenness centrality of above 1500: The United States (4222.4), Russia (3349.3), France (1981.5) and Canada (1591.4). Most other countries are much more peripheral in the alliance network. This observation is consistent with the view of the so-called hub-and-spoke model of international relations described by Victor Cha [5], where in the post-World War II world, several key countries play the most important role on the global stage, and all other countries have their relationship defined by their connection to one or more of these "hubs."

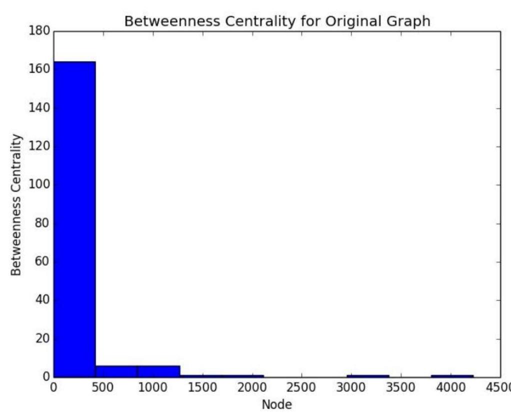

Fig. 3

It is of interest to note that the overlap in the neighbourhoods of these four hub countries vary quite widely. Fig. 4 summarises this. Russia shares only 19 and 16 countries in its immediate alliance neighbours with the US and Canada respectively, while the US and Canada share as many as 58 such countries. While this may at first suggest that the world's alliance network is divided into a set of subgraphs clustered around hub countries like the US and Russia, Table 1 suggests that the composition of the global alliance network is more nuanced than this multipolar view of international relations. France shares almost as many neighbours with Russia as with Canada, and not substantially more than with the US. A key point to note about the hub-and-spoke model is that countries at the 'periphery' can be connected to multiple hubs. While Fig. 4 highlights the polarisation of the world between the hubs in Russia and North America, it also suggests that there is a sliding scale in the shared neighbourhoods of this network.

	USA	Canada	France
Canada	58		
France	32	26	
Russia	19	16	25

Fig. 4

Fig. 5 is the distribution of the betweenness centrality of our false positive graph. We notice that, although the degree distribution of the nodes in this graph were very different from the original graph of extant alliances, the distribution of betweenness centrality is remarkably similar. This may suggest that similar dynamics of the hub-and-spoke model of international relations are at play.

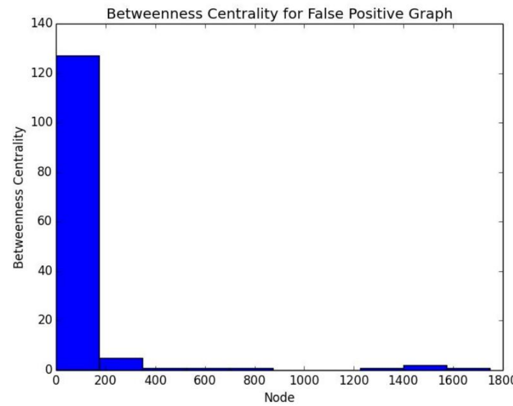

Fig. 5

The implications for the hub-and-spoke model in the false positive network must be different for the map of already extant alliances, however, since this network explicitly exclude all current alliances. The centrality of the handful of countries with a lot of influence in Fig. 5 can be interpreted as their potential in creating great changes in the global alliance network.

Three countries have a betweenness centrality in the shadow alliance network of above 1500, Australia (1748.9), New Zealand (1551.1) and the Philippines (1534.2). Our previous observations with regard to the high degree of the Philippines can be extended here. All three countries have export-oriented economies with Christian majorities, and strong former colonial ties with the Anglophone and Euro-North American World. However, all three countries sit at the far edge of the Pacific Rim, where the geopolitical forces of the Cold War and other historical events were less pressing than in the North Atlantic, because of their sheer geographical distance away from the other countries. This helps explain why these three countries have the greatest unrealised potential influence in the global alliance network.

Nevertheless, because of their centrality in the false positive network, we evaluate them as having the greatest influence in changing the formal shape of the global alliance network. These three countries are the nodes that are not only best positioned to form immediate pairwise alliances with other countries, but are also best positioned to shift entire subgraphs of our 'shadow alliance' false positive network into the real-world graph in the future.

Conclusion

Our results indicate that country contiguity, past disputes, dominant religion, and trade are strong predictors of alliance formation, achieving a 91% prediction accuracy. Analysis of the graph structure constructed with false positives generated by our prediction system suggest a shadow alliance network with a similar hub-and-spoke structure to the real-world alliance network. We can infer that our prediction system preserves and captures the hub-and-spoke model of global alliances because of the similar distributions of betweenness centrality.

Future work

Further work can be done to examine which features have the most impact on our prediction results. Taking a slightly different approach to global alliances, one can explore the feature space by using methods of variable selection with stepwise approaches and best subset, then evaluate and compare different sets of variables to measure the impact of each. Another interesting approach would be to preprocess the data by following the "clustering" algorithm described by Nowell and Kleinberg to clean up the data [4]. The procedure involves clustering nodes in our graph, removing tenuous edges, and running our predictor on the resulting graph. An example they describe is computing a *score(u, v)* for all the original edges and removing a certain proportion of lowest score edges. This score can be computed with a combination of the features. The intuition behind this approach is to be able to run the predictor only important edges and to reduce noise. It is also viable to explore more methods of cross-validation - leave one out (LOO), k-fold, random permutations. This can help with improving the training of the model, as well as potentially avoiding overtraining or bias.

Finally, another interesting analysis of the graph would be to use the disputes dataset to introduce negative signed edges to the graph. This would allow for signed edge and triad analysis, as well as predicting signed links.

Bibliography:

[1] Jackson, Matthew and Stephen Nei. "Networks of military alliances, wars, and international trade." In *Proceedings of the National Academy of Sciences of America* 112, no.50 *(2015*): 15277–15284.

[2] Guha, Ramanthan, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. "Propagation of trust and distrust." In *Proceedings of the 13th international conference on World Wide Web*, pp. 403-412. ACM, 2004.

[3] "About The Correlates of War Project" Last modified 2016. http://www.correlatesofwar.org/.

[4] Liben-Nowell, David, and Jon Kleinberg. "The link-prediction problem for social networks." *Journal of the American society for information science and technology* 58, no. 7 (2007): 1019-1031.

Datasets

[1] Gibler, Douglas M. 2009. International military alliances, 1648-2008. CQ Press.

[2] Correlates of War Project. *Direct Contiguity Data, 1816-2006*. Version 3.1. Online: http://correlatesofwar.org.

[3] Barbieri, Katherine and Omar Keshk. 2012. Correlates of War Project Trade Data Set Codebook, Version 3.0. Online: http://correlatesofwar.org.

[4] Zeev Maoz and Errol A. Henderson. 2013. “The World Religion Dataset, 1945-2010: Logic, Estimates, and Trends.” International Interactions, 39: 265-291.
[5] Cha, Victor D. "Powerplay: Origins of the US alliance system in Asia." *International Security* 34.3 (2010): 158-196.