Insights in International Relations

Submitted as Term Project for MATH E-23C

Mohit Negi, Ferran Vega

Spring 2020

Abstract

The primary dataset used in this project is Gary King's "10 Million International Dyadic Events" dataset from Harvard dataverse. The data available here include almost 10 million individual events, each coded to the exact day they occur or become known. Each event is summarized in the data as "Actor A does something to Actor B", with Actors A and B recording about 450 countries and other (within-country) actors and "does something to" coded in an ontology of about 200 types of actions. The data are coded by computer from millions of Reuters news reports. In some topics, we filter this huge dataset to only include "Inter-country" events. In addition to this, we use several other datasets that are cited appropriately in the relevant topics. The idea of the project is to utilize the extremely rich dataset to find insights in International Relations. For this, we employ various data analysis techniques.

Contents

1 : Peer Influence in International Relations.

Abstract :

In this topic, we exploit the properties of the Exponential and Weibull probability distributions to see which Inter-country dyadic events might be peer-influenced and which might not.
The Hazard rate (aka Failure rate) is the conditional probability of "failure" on survival upto a point in time. For a random variable T, it is defined as : $\frac{f_T}{1-F_T}$, where f_T is the PDF and F_T is the CDF of T.
The Exponential distribution (as we'll see in a bit) has a constant Hazard rate whereas the Weibull with a shape parameter $(\beta) < 1$ has decreasing Hazard rate.
This property makes the Exponential distribution exhibit "Memoryless-ness" and the Weibull exhibit "Infant-mortality".
We use these properties to explore an interesting insight : Some types of events seem to be peer-influenced while some don't.

In Part I, we discuss our analysis strategy.
In Part II, we look at an example of peer influenced events.
In Part III, we present a counter-example that does not show peer influence.
In Part IV, we discuss the results and point out the caveats. Datasets used :

- 10 Million International Dyadic Events (10MM_IDE)

I. Strategy

We begin by looking at the Poisson process. If the number of events occuring in unit time is a random variable $X \sim Pois(\lambda)$, taking a time interval $[0, \infty)$, the number of events occuring in every subinterval of length t will be a random variable with distribution $Pois(\lambda t)$. Further, the number of events occuring in distinct arbitrary non-overlapping intervals are independent random variables.
This is a Poisson process.

If we look at the time intervals T between successive occurences in such a Poisson process, they are random variables $T \sim Expo(\lambda)$. This random variable T exhibits the property of "Memoryless-ness" defined as : $P(T > t + x | T = x) = P(T > t)$.
It is also provable that if X is a positive, continuous random variable that is memoryless, then $\exists \lambda \in \mathbb{R}$ such that $X \sim Expo(\lambda)$.

This is useful to see if the occurence of some event in the world is purely random. Our primary dataset 10MM_IDE includes the date of every event of a certain type between an "Actor" and a "Target". We filter the data to include only those events where both the parties are of the type "Country".

The strategy is to determine if the time interval T between successive events of a certain type is exponentially distributed. If it is, it implies that the event occurs purely at random and is not "peer influenced".

By Peer influence, we mean that the probability of an event occuring is dependent on the occurence of the event before it. If an event occurs between a dyad, it "influences" the probability of the event occuring between other dyads.

It is important to note, however, that if T is not exponentially distributed, it need not imply that the event is peer influenced. For this, we need the Weibull distribution.

The Weibull distribution is a generalization of the Exponential distribution and has the PDF :

$$f_T = \frac{\beta t^{\beta-1}}{\eta^\beta} e^{-(\frac{t}{\eta})^\beta}$$

and CDF :

$$F_T = 1 - e^{-(\frac{t}{\eta})^\beta}$$

The Hazard rate is :

$$H_T = \frac{f_T}{1 - F_T} = \frac{\beta}{\eta}(\frac{t}{\eta})^{\beta-1}$$

where β is the shape parameter and η is the scale parameter.
It is easy to see that when $\beta = 1$, the functions describe the Exponential random variable.

We make the functions in R,

```
#Weibull PDF function.
weibull_pdf <- function(t,eta,beta){
  beta*(t)^(beta-1)*exp(-1*(t/eta)^beta)/(eta^beta)
}

#Integrating to find the CDF.
weibull_cdf <- Vectorize(function(p,eta,beta){
  integrate(weibull_pdf,lower=10^(-100),upper=p,eta=eta,beta=beta)$value
})

#Hazard rate will be PDF/survival.
hazardrate <- Vectorize(function(x,eta,beta){
  weibull_pdf(x,eta,beta)/(1-weibull_cdf(x,eta,beta))
})
```

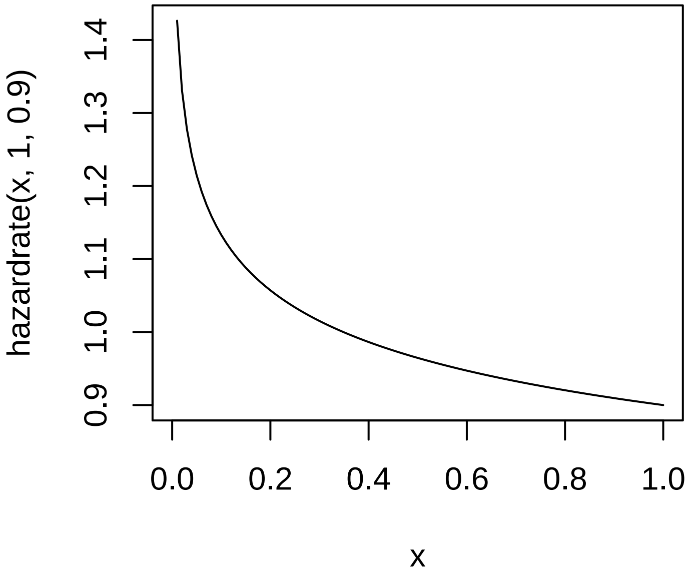
It is interesting to note what happens to the H_T as time progresses.
For this, we differentiate H_T w.r.t. t.

$$H_T' = \frac{\beta}{\eta^\beta}(\beta - 1)t^{\beta-2}$$

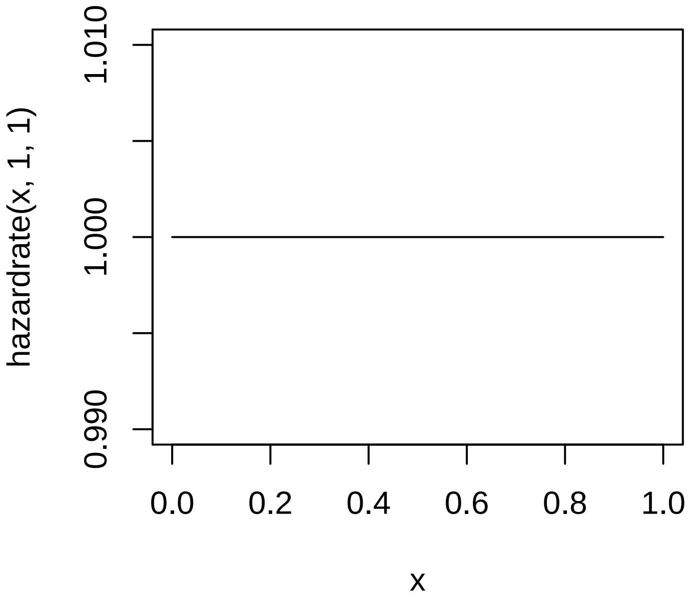
- When $\beta < 1$, H_T is decreasing. : The probability of the event occuringdecreases over time. Every additional period of survival implies a longer remaining life expectancy (when event occuring is death). This is called the "Lindy Effect". This property is likened to "Infant Mortality".
- When $\beta = 1$, H_T is constant. : This is what we call "Memoryless-ness".
- When $\beta > 1$, H_T is increasing. : The probability of the event occuring increases over time. This is likened to the "wear-and-tear" effect on machine failure events.

Let's verify this :

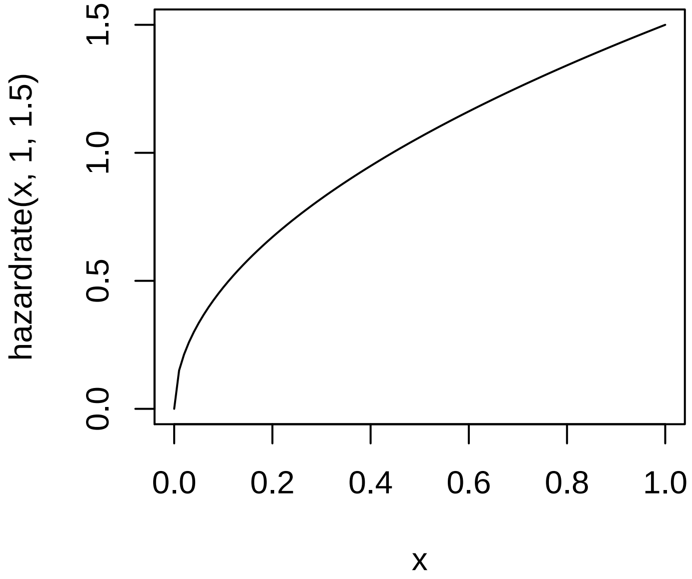
```
curve(hazardrate(x,1,0.9)) # Weibull with beta < 1 has decreasing hazard rate.
```



```
curve(hazardrate(x,1,1)) # Weibull with beta = 1 has constant hazard rate.
```



```r
curve(hazardrate(x,1,1.5)) # Weibull with beta > 1 has increasing hazard rate.
```



For our purposes, a Weibull r.v. with $\beta < 1$ can imply a "positive" peer influence. The event encourages more events. When $\beta > 1$, we see "negative" peer influence. The event discourages subsequent events which become more probable as time passes since the last event.

II. Taking the theory to data (Armed Assistance)

- Armed Assistance Requests

First we look at the events of the type “Ask for armed assistance”. Let’s see if it follows any of our interesting distributions.

We load the data in, do some cleaning and then create a vector that holds the time intervals between subsequent events through the following for loop.

```
for(i in 1:(nrow(data)-1)){
  if(SrcName[i] != SrcName[i+1] & SrcName[i] != TgtName[i+1]
     & TgtName[i] != TgtName[i+1] & TgtName[i] != SrcName[i+1]){
    j <- difftime(strptime(data$EventDate[i+1], format = "%Y-%m-%d"),
                  strptime(data$EventDate[i], format = "%Y-%m-%d"),units=timeframe)
    vec90 <- c(vec90,j)
  }
}
detach(data)
```

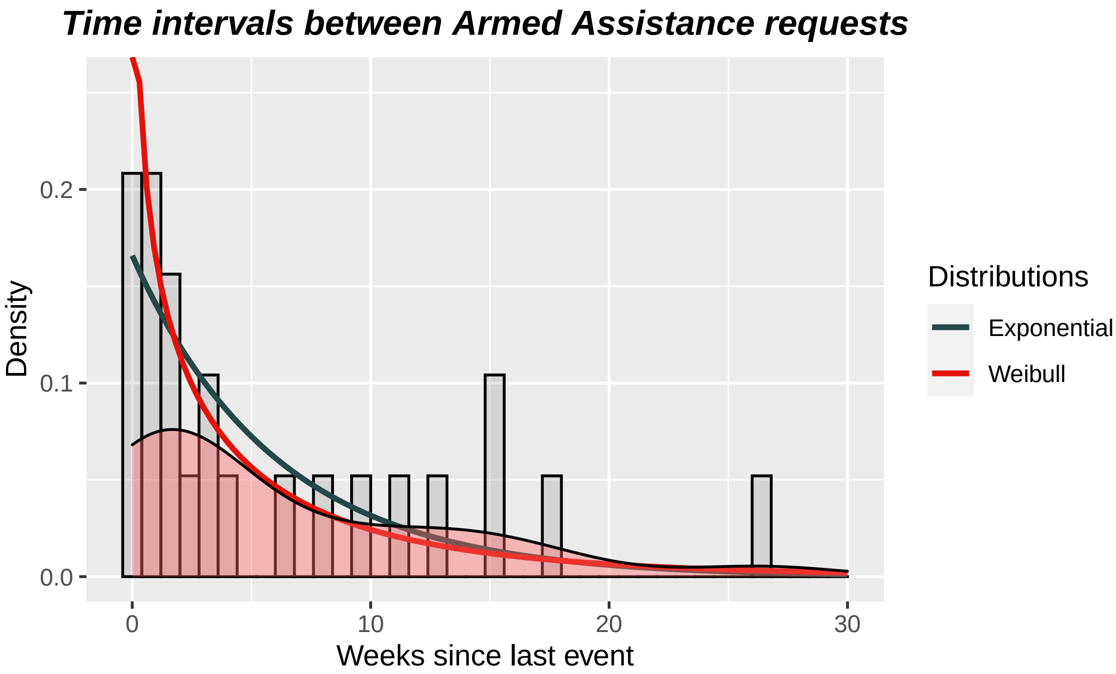
This for loop populates the vec90 vector with date differences of successive events.
It is computed such that none of the actors(src or tgt) are repeated in successive events.
This is done to avoid mass-action events like a country asking for armed assisstance from multiple countries at once or multiple countries asking a single country for armed assisstance at once.
Such events would obviously be peer influenced.
By doing this, we remove such trivial examples of peer-influence.

We use Likelihood Ratio test for g.o.f of exponential and weibull distributions.

```
    Likelihood ratio test for the Exponential distribution

data:  vec90
S = 2.9502, p-value = 0.09
sample estimates:
[1] 0.165844

    Test based on the Exponentiated Weibull distribution for the Weibull
    distribution using the MLEs with the Likelihood ratio procedure

data:  vec90
S = 1.0144e-06, p-value = 0.995
sample estimates:
      eta      beta
5.1813165 0.7685772
```

The weibull fits excellently (pval > 0.9) and estimates a shape parameter $\neq 1$. This gives sufficient evidence that armed assisstance requests are NOT purely random and they might be peer influenced. The weibull shape parameter is 0.769 which tells us there is positive peer influence.

Let’s graph our fitted curves.



Let’s calculate the Expected number of weeks between two subsequent armed assistance requests.

```
integrand1 <- function(t,eta,beta){
  t*(beta*(t)^(beta-1)*exp(-1*(t/eta)^beta)/(eta^beta))
}

exp_days <- function(eta,beta){
  integrate(integrand1,lower = 10^(-100),
            upper = Inf,eta=eta,beta=beta)$value}

exp_days(w1$estimate[1],w1$estimate[2])
```

```
[1] 6.049117
```

Which tells us that the Expected number of weeks between two subsequent armed assistance requests is about 6 weeks.

III. Taking the theory to data (Ultimatums)

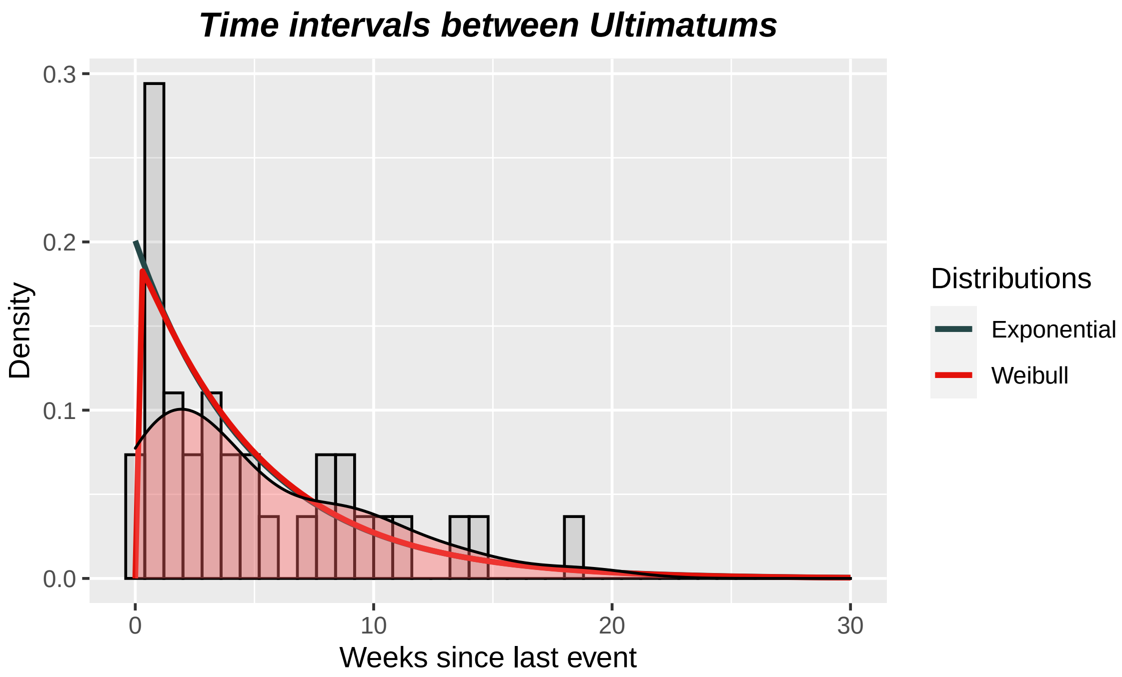
- Giving Ultimatums

Now we look at the events of the type "Give ultimatum". Let's see if it follows any of our interesting distributions.

In a similar manner, we generate the vector and conduct a Likelihood Ratio test on it.

```
    Likelihood ratio test for the Exponential distribution

data:  vec90
S = 0.014257, p-value = 0.93
sample estimates:
[1] 0.2006745

    Test based on the Exponentiated Weibull distribution for the Weibull
    distribution using the MLEs with the Likelihood ratio procedure

data:  vec90
S = 8.1441e-06, p-value = 0.97
sample estimates:
     eta     beta
5.017213 1.016477
```

The exponential fits well and the weibull fits excellently (pval > 0.95) and estimates a shape parameter ≈ 1. This gives strong evidence that giving ultimatums might be purely random events. The weibull shape parameter is 1.02. We can't say it is exactly exponential but for such a small data sample, it's a good approximation.



IV. Taking the theory to data (Discussion)

- Discussion

We can draw two conclusions from our analysis. Firstly, we see that the act of asking other countries for military aid might exhibit peer influence.
The act is defined in such a way that it includes requests of “peacekeeping” forces to international agencies like the UN.
We filtered the data in such a way that each event was “distinct” than its predecessor in the sense that they didn’t have either party in common.
A plausible theory might be that during war-time, each belligerant might call upon its allies at approximately the same time.

Secondly, we saw that the act of giving ultimatums might be purely randomly distributed in time. This is reasonable as there doesn’t seem to be any reasons for multiple distinct dyads to influence each other’s actions.

It is also important to note the caveats:

- Due to small sample sizes, the reliability of the results is not perfectly solid.
- We only took data from 1990-1995 as taking the aggregate data (1990-2005) did not yield the same results. This might be due to conditions changing between the long time period that changes λ. For example, if due to some change in global conditions, the average number of ultimatums in a given week increases after 1995, it would no longer be a Poisson process with a unique λ. This means if we aggregate pre and post 1995 data, we won’t see the memoryless property even if it exists within the separate time periods.

2 : Does providing Aid buy Softpower?

Abstract :

It is widely believed that the provision of Economic Aid by countries to those in need is motivated by more than just good faith. It is common to see countries being accused of “buying favor” through their charitable acts. The most famous example of this might be the “Belt and Road Initiative (BRI)” that was launched by the Chinese government in 2013.
Many skeptics have raised concerns over the project being a form of “Neo-colonialism”. A favourite example towards this is the Chinese takeover of the Hambantota Port in Sri Lanka for use as a Naval base in the Indian Ocean when the latter was unable to repay debts to China.
If this phenomenon is indeed true, it may be beneficial to understand if it is effective.
For this, we analyze the 10MM_IDE dataset, looking at the data on Inter-country dyadic events of the form “Provide Economic Aid” and “Threaten”. We use these variables as proxies for “Aid” and “Softpower”.

The idea is to see if providing Aid to a country significantly impacts the probability of receiving a Threat from it. If this impact turns out to be negative, we can conclude that providing Aid indeed buys Softpower.

In Part I, we conduct some preliminary data analysis to motivate the use of Network Analysis techniques.
In Part II, we actually go about analyzing the the Aid and Threat networks from 1990-2005 through Exponential Random Graph Modelling.
In Part III, we interpret the results and conclude the topic.

Datasets used :

- 10 Million International Dyadic Events (10MM_IDE)
- Correlates of War Dataset
- HDI dataset, UNDP
- GDP per capita dataset, World Bank

I. Preliminary Analysis

Here, we utilize a similar strategy as the last topic wherein we explained how an Exponential distribution can be used to determine if events of a certain type occur purely at random.
When looking at Network graphs where the different countries are represented as Nodes and dyadic events are represented as the Edges, we can determine the probability of edge formation through Network analysis techniques. This probability p can either be constant across the entire graph or be determined by features such as nodal and edge attributes, neighbouring edges, etc. A graph of the former type is known as an “Erdos-Renyi Random Graph”.

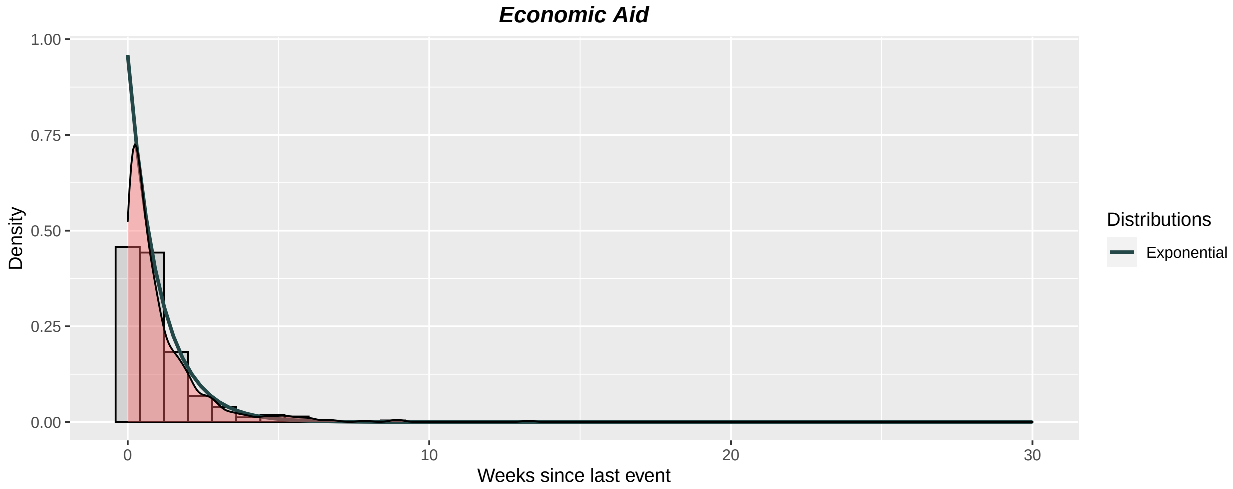
If find an event to occur purely at random, we should expect to see an Erdos-Renyi Graph. If not, we might find the graph to be more complex. In such a graph, ties would not be formed at random but might be dependent on certain nodal and edge attributes. We model such dependencies by using the “Exponential Random Graph Model (ERGM)”.
With this in mind, we try to provide motivations for the use of an ERGM in this section.
We generate the vectors containing time intervals between successive Aid and Threat events using a similar loop as in the last topic. We conduct Likelihood Ratio tests on both these events to check for goodness of fit of an Exponential distribution.

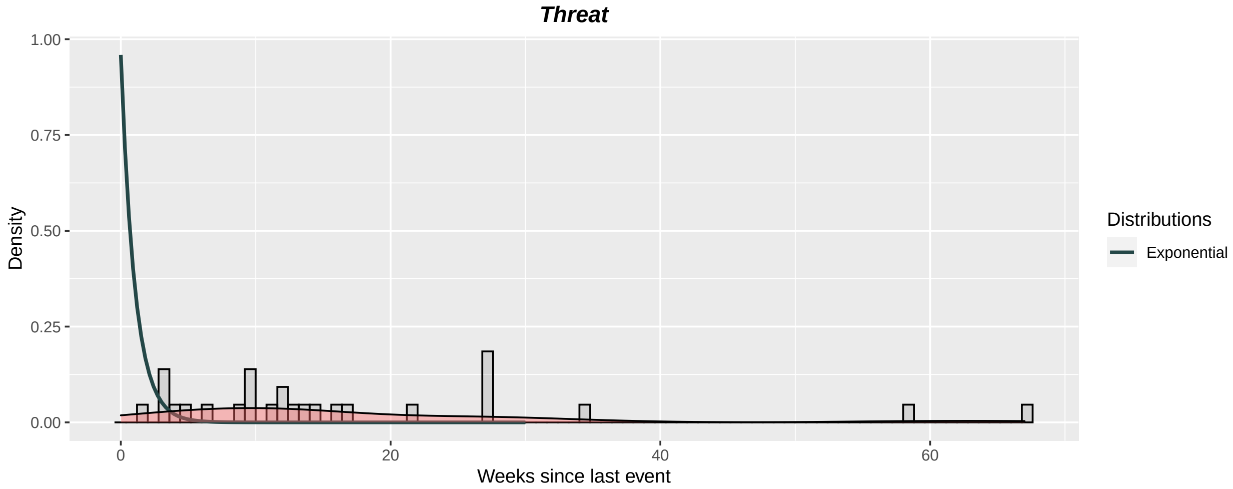
We graph these fits.

```
    Likelihood ratio test for the Exponential distribution

data:  aid_totalvec
S = 640.6, p-value < 2.2e-16
```

```
sample estimates:
[1] 0.9589257

    Likelihood ratio test for the Exponential distribution

data:  threat_totalvec
S = 1.6508, p-value = 0.23
sample estimates:
[1] 0.05820758
```

We can clearly see that the Exponential does NOT fit either Aid or Threat data well (pvals ≈ 0). This gives sufficient evidence that both the events are NOT purely random. Edge formation might be dependent on other features.





This result motivates us to conduct an ERGM to analyze these networks as they are clearly not Erdos-Renyi Random Graphs.

II. Network Analysis using ERGMs

The Exponential family is a broad family of models for covering many types of data, not just networks. An Exponential Random Graph Model is a model from this family which describes networks.

Formally a random graph Y consists of a set of n nodes and m dyads (edges) $Y_{ij} : i = 1, 2, 3...n; j = 1, 2, 3...n$ where $Y_{ij} = 1$ if the nodes (i, j) are connected and 0 otherwise.

The basic assumption of these models is that the structure in an observed graph y can be explained by any statistics $s(y)$ depending on the observed network and nodal attributes.

The model is defined as :

$$P(Y = y|\theta) = \frac{exp(\theta^T s(y))}{c(\theta)}$$

where,
θ is a vector of model parameters associated with $s(y)$,
$c(\theta)$ is the normalizing constant to make it a probabiliy measure between 0 and 1. It is equal to $\sum_{y \in Y} exp(\theta^T s(y))$

These models represent a probability distribution on each possible network on n nodes and may be directed or undirected.

We can interpret the results of such models in the following way :

We define Y_{ij}^c as the graph we observe before edge Y_{ij} is measured. So $P(Y_{ij} = 1|Y_{ij}^c)$ is the probability of formation of an edge between i and j given the rest of the graph.

$$\begin{aligned} odds(Y_{ij} = 1) &= \frac{P(Y_{ij} = 1|Y_{ij}^c)}{1 - P(Y_{ij} = 1|Y_{ij}^c)} \\ &= \frac{P(Y_{ij} = 1|Y_{ij}^c)}{P(Y_{ij} = 0|Y_{ij}^c)} \\ &= \frac{exp(\theta^T s(Y_{ij}^+))}{exp(\theta^T s(Y_{ij}^-))} \end{aligned}$$

Here, Y_{ij}^+ is the graph with edge Y_{ij} present and Y_{ij}^- is the graph where its absent. It is important to note that both these graphs are identical $\forall Y_{pq} \neq Y_{ij}$.

$$\begin{aligned} odds(Y_{ij} = 1) &= exp(\theta^T s(Y_{ij}^+) - \theta^T s(Y_{ij}^-)) \\ &= exp(\theta^T [s(Y_{ij}^+) - s(Y_{ij}^-)]) \end{aligned}$$

Here, $s(Y_{ij}^+) - s(Y_{ij}^-)$ is the “change statistic” vector that represents the change in the values of the statistics by adding the edge Y_{ij}. Each variable now accounts for the change in counts of network statistics. We notate it as $s(\Delta Y_{ij})$.

$$\begin{aligned} odds(Y_{ij} = 1) &= exp(\theta^T s(\Delta Y_{ij})) \\ logodds(Y_{ij} = 1) &= \theta^T s(\Delta Y_{ij}) \\ logit(Y_{ij} = 1) &= \theta_1 s_1(\Delta Y_{ij}) + \theta_2 s_2(\Delta Y_{ij}) + \theta_3 s_3(\Delta Y_{ij}) + ...\theta_k s_k(\Delta Y_{ij}) \end{aligned}$$

This is the Logit model we solve when we use an ERGM. The statistics can be network features like number of edges, triangles, etc.

Now suppose we choose only “Edges” as the network statistic.
$\theta_1 s_1(\Delta Y_{ij}) = \theta_1(1) = \theta_1$
This is because the change in the edge count by adding edge Y_{ij} is 1.

Our model thus becomes,

$$\begin{aligned} logodds(Y_{ij} = 1) &= \theta_1 \\ odds(Y_{ij} = 1) &= exp(\theta_1) \\ \frac{p}{1-p} &= exp(\theta_1) \\ p &= \frac{exp(\theta_1)}{1 + exp(\theta_1)} \end{aligned}$$

This p would be the “Erdos-Renyi” probability of an edge forming randomly. However, the reason we use ERGMs is to model complex networks where p is not constant across the entire network. Just like any other Logit model, we control for these other factors by adding them as covariates in our Logit model.

For example,

$$logodds(Y_{ij} = 1) = \theta_1 Edges + \theta_2 GDP_i + \theta_3 GDP_j + \theta_4 Border$$

We cleaned the following datasets to obtain our covariates :

1. GDP per capita
2. Dominant Ethnic Group
3. HDI
4. Borders
5. Alliances : Correlates of War dataset
6. Wars : Correlates of War dataset

We ruled out the following due to issues : 1. Dominant Ethnic Group : Data was available only for a small subset of countries which would pose problems due to the network being sparse and possibility of selection bias. 2. Wars : ERGM models did not converge possibly due to very high collinearity. We suspected the War and alliance data might be too correlated.

The strategy here is to specify “Economic Aid” as a directed edge covariate that is equal to 1 if country i received Economic Aid from country j and 0 otherwise. Y_{ij} is a directed edge from i to j representing a “Threat” made by country i to j. A negative θ on the predictor variable “Aid” will mean that receiving aid from a country j reduces the probability of sending a threat to that same country for all countries i. We explore caveats regarding direction of causality in the next section.

We used an extension of the ERGM called the Temporal ERGM (TERGM) to analyze an evolving network where the nodal and edge covariates keep changing over time. We use data from 1990-2005.

For every year, we used a loop to extract all the data and store it into a Network object created using the “btergm” package in R.

```
  threatnet <- network(threat_adj_1990$adjacency,directed = TRUE,matrix.type = "adjacency")

  network::set.vertex.attribute(threatnet, 'Per Capita Income', as.numeric(node.att.1990$GDP))
  network::set.network.attribute(threatnet,'Alliance', edge.att.1990$Alliance)
  network::set.network.attribute(threatnet,'Border', edge.att.1990$border)
  network::set.network.attribute(threatnet,'Aid', edge.att.1990$Aid)
```

We then run the model with the following specification,

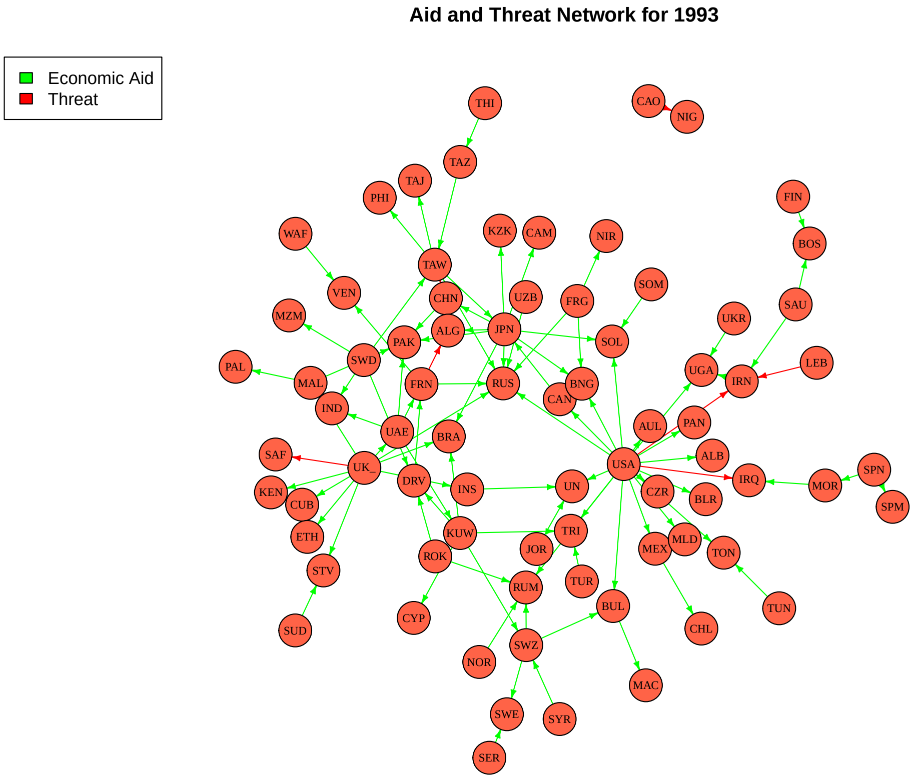
```
model_fulltime11 <- btergm(netlist ~ edges + mutual()
                           + nodeocov('Per Capita Income')
                           + nodeicov('Per Capita Income')
                           + edgecov('Alliance')
```

```
                          + edgecov('Border')
                          + edgecov('Aid'),
                          R = 50)
summary(model_fulltime11)
```

We obtain the results,

```
                                Estimate          2.5%         97.5%
edges                       -9.302635466 -10.314917561  -8.739616494
mutual                     -12.999570074 -14.738050746 -11.969415536
nodeocov.Per Capita Income  -0.002061352  -0.008039130   0.002551004
nodeicov.Per Capita Income  -0.003813694  -0.009213134   0.003991090
edgecov.Alliance             1.014185541  -0.171276836   1.897751214
edgecov.Border              -0.031985425 -15.421425935   1.296193712
edgecov.Aid                -11.906023646 -13.825718117 -10.955198842
```

III. Discussion



The sparse nature of the network makes it very difficult to effectively interpret the marginal effects of our predictor variables. The edges term has a coefficient of -9.511 which translates to, holding all others terms to be 0, a base probability of 0.007% of an edge ("Threat") forming randomly. This is a clear indication of a very sparse network which is reasonable since threats are not that common when considering all the countries in the world. The mutual term, which measures reciprocity in edges, has a very low coefficient of -13.853 which tells us reciprocity was very rarely observed.

Since the probability to logodds transformation is monotonic, a negative coefficient of -11.906 on "Aid" can be said to imply that providing economic aid to a country reduces the probability that it will threaten you (it reduces the odds of a threat event by $1 - e^{-11.906} \approx 99.99933\%$), which is consistent with our main hypothesis.

However, these estimates might just be a result of a really sparse network with too few threat events.

3 : The (not-so)Great Escape : Analyzing Political Exile.

Abstract :

In January 1979, after 8 years of despotic rule, Ugandan dictator Idi Amin found himself surrounded by the Tanzanian Army, which had launched a counter-offensive against his forces. As the Ugandan Army retreated, Amin fled into exile to Libya, then ruled by Muammar Gaddafi.
In this topic, we try to explore what factors might determine where dictators (and politicians in general) flee to when escaping crises. We employ the same Network Analysis framework to analyze the "Political Flight" network constructed from the 10MM_IDE dataset. We try to find the attributes, both nodal and edge-wise, that might have a significant impact on the probability of observing an inter-country political exile edge.

In Part I, we conduct some preliminary data analysis to motivate the use of Network Analysis techniques.
In Part II, we actually go about analyzing the the Aid and Threat networks from 1990-2005 through Exponential Random Graph Modelling.
In Part III, we interpret the results and conclude the topic.

Datasets used :

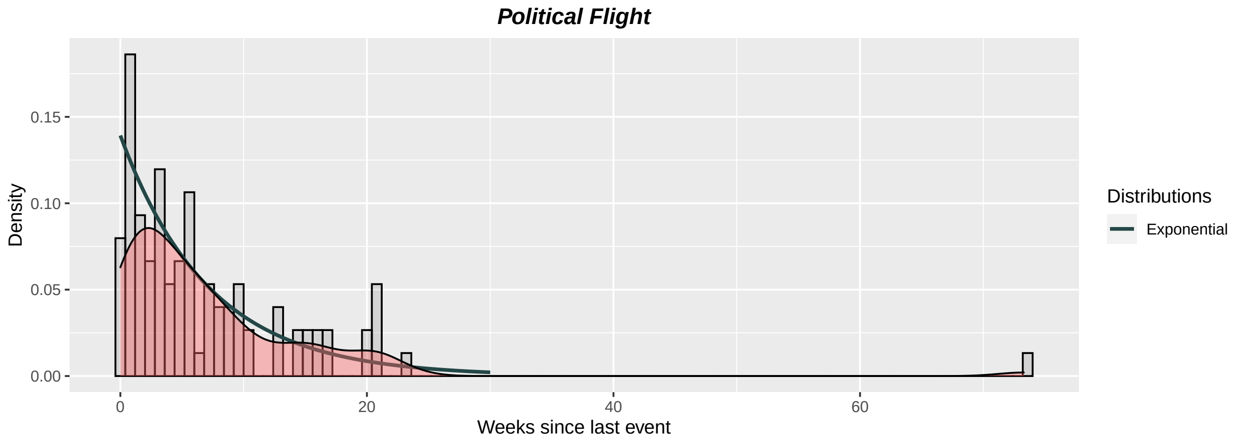
- 10 Million International Dyadic Events (10MM_IDE)
- GDP per capita dataset, World Bank
- Correlates of War dataset
- HDI dataset, UNDP
- Religious composition by country, Pew Research Center

I. Preliminary Analysis

We follow the exact procedure as the last topic to see if data on events of the type "Political Flight" follow the Exponential distribution. We'll use "Asylum" as a short-hand for "Political Flight" in this paper.

```
    Likelihood ratio test for the Exponential distribution

data:  asylum_totalvec
S = 35.012, p-value < 2.2e-16
sample estimates:
[1] 0.1392298
```

We can clearly see that the Exponential does NOT fit Asylum data well (pval ≈ 0). This gives sufficient evidence that the events are NOT purely random. Edge formation might be dependent on other features.



This result motivates us to conduct an ERGM to analyze these networks as they are clearly not Erdos-Renyi Random Graphs.

II. Network Analysis using ERGMs

We employ the same procedure as the last topic. Here we specify Y_{ij} as a directed edge from i to j representing a politician fleeing from i to j.

We cleaned the following datasets to obtain our covariates :

1. GDP per capita
2. HDI
3. Dominant Religious Group
4. Civil conflicts : 10MM_IDE
5. Borders
6. Alliances : Correlates of War dataset
7. Wars : Correlates of War dataset

We ruled out the following due to issues : 1. Ethnic Homogeneity : Data was available only for a small subset of countries which would pose problems due to the network being sparse. 2. Wars : ERGM models did not converge possibly due to very high collinearity. We suspected the War and alliance data might be too correlated.

We use a TERGM like before to model the evolving network.

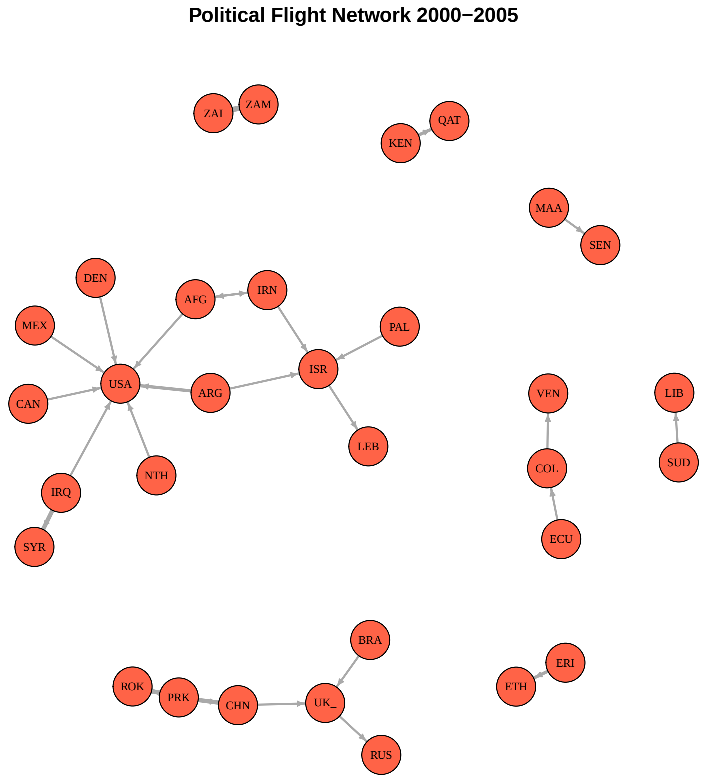
We then run the model with the following specification,

```
asylfull6_rel <- btergm(netlist ~ edges + mutual()
                           + nodeocov('Per Capita Income')
                           + nodeicov('Per Capita Income')
                           + nodematch('Religion')
                           + nodeocov('Civil Conflicts')
                           + nodeicov('Civil Conflicts')
                           + nodeocov('HDI')
                           + nodeicov('HDI')
                           + edgecov('Alliance')
                           + edgecov('Border')
                           ,
                           R = 50)
```

We obtain the results,

```
                                Estimate          2.5%        97.5%
edges                      -9.554403e+00 -13.987505139 -7.713146483
mutual                     -1.121583e+01 -13.644752987 -9.149985882
nodeocov.Per Capita Income -2.349755e-04  -0.002711913  0.003555899
nodeicov.Per Capita Income -2.271146e-03  -0.006878858  0.002195479
nodematch.Religion          4.475696e-01   0.108754390  0.951856649
nodeocov.Civil Conflicts    8.606049e-02   0.056021072  0.111564722
nodeicov.Civil Conflicts    7.560511e-02   0.044126878  0.161976805
nodeocov.HDI                3.832585e-01  -1.985858187  2.211232378
nodeicov.HDI                1.012813e+00  -1.494271806  5.073881709
edgecov.Alliance            5.336432e-01  -1.031423457  1.495390309
edgecov.Border              1.659626e-02 -10.526675782  1.613296451
```

III. Discussion



Judging from the extremely low coefficient of -9.554 on the edges term, we observe a sparse network ($p \approx 0.007\%$). This is reasonable since political flight events are not that common. The mutual term, which measures reciprocity in edges, has a very low coefficient of -11.215 which tells us reciprocity was very rarely observed.

The term "nodematch.Religion" measures homophily in the network with respect to "major religion". It has a positive coefficient of 0.447. It tells us that countries having the same major religion are more likely to form an edge, which in our case means having the same major religion increases odds of a politician fleeing to that country by about $e^{0.447} - 1 \approx 56\%$ when compared to countries with different major religions. This might be reasonable as religion can plausibly impact cultural homogeneity and political alliances. It is important to note that the Alliance term represents "Defence Alliances" only.

Civil conflicts in host as well as origin countries seem to positively impact the probability of a political flight. The term nodeocov.Civil Conflicts represents civil conflicts in origin country and increases odds of a political flight by $e^{0.0861} - 1 \approx 8.9\%$ which seems resonable as countries with internal turmoil might be more likely to see its politicians fleeing. The term nodeicov.Civil Conflicts represents civil conflicts in host country and increases odds of a political flight by $e^{0.0756} - 1 \approx 7.85\%$ which might be due to politicians fleeing to similarly "unstable" regimes as was the case in our introductory example of Idi Amin fleeing to Libya.

4. Minding their own business : Does Geographical Isolation impact development?

Abstract : In this short topic, we analyze HDI and Geographical contiguity data to determine if there is a statistically significant impact of having no borders i.e. being isolated on HDI levels. For this, we make use of the Pearson's Chi-squared test. We then use a Permutation test after which we compare the two methods.

Datasets used :

- HDI dataset, UNDP

The UNDP classifies each country into one of three development groups: Low human development for HDI scores between 0.0 and 0.5, Medium human development for HDI scores between 0.5 and 0.8. High human development for HDI scores between 0.8 and 1.0. We convert our HDI data to these categorical variables.

We conduct our analysis using the following dataset,

```
  country   HDI No.border HDI_cat
1     AFG 0.387         0     Low
2     ALB 0.687         0  Middle
3     ALG 0.676         0  Middle
4     AND 0.820         0    High
5     ANG 0.428         0     Low
7     ARG 0.775         0  Middle
```

This is the contingency table,

```
         HDI_cat
No.border Low Middle High
        0  39     70   28
        1   3     18    7
```

The Chi square test used in the Contingency table approach requires at least 80% of the cells to have an expected count greater than 5 or else the sum of the cell Chi squares will not have a Chi square distribution. In our case, only 1/6 of the cells have a count lower than 5.

We conduct the Chi-squared test.

```r
ChiSq <-function(Obs,Exp){
  sum((Obs-Exp)^2/Exp)
}

chisq <- ChiSq(Obs,Exp) ; chisq
```

```
[1] 3.865075
```

```r
pvalue <- pchisq(chisq, dfs, lower.tail = FALSE); pvalue
```

```
[1] 0.1447803
```

This pvalue is not small enough for us to reject the null hypothesis of independence. There is over a 14% chance that the observed data pattern could have arisen by chance under the assumption that the null hypothesis of independence is true. We are not comfortable with such a high probability for making a type I error (false positive), and thus conclude that the we cannot reject the null hypothesis of independence in this case.

```
observed <- Mean.Border - Mean.NoBorder ; observed
```

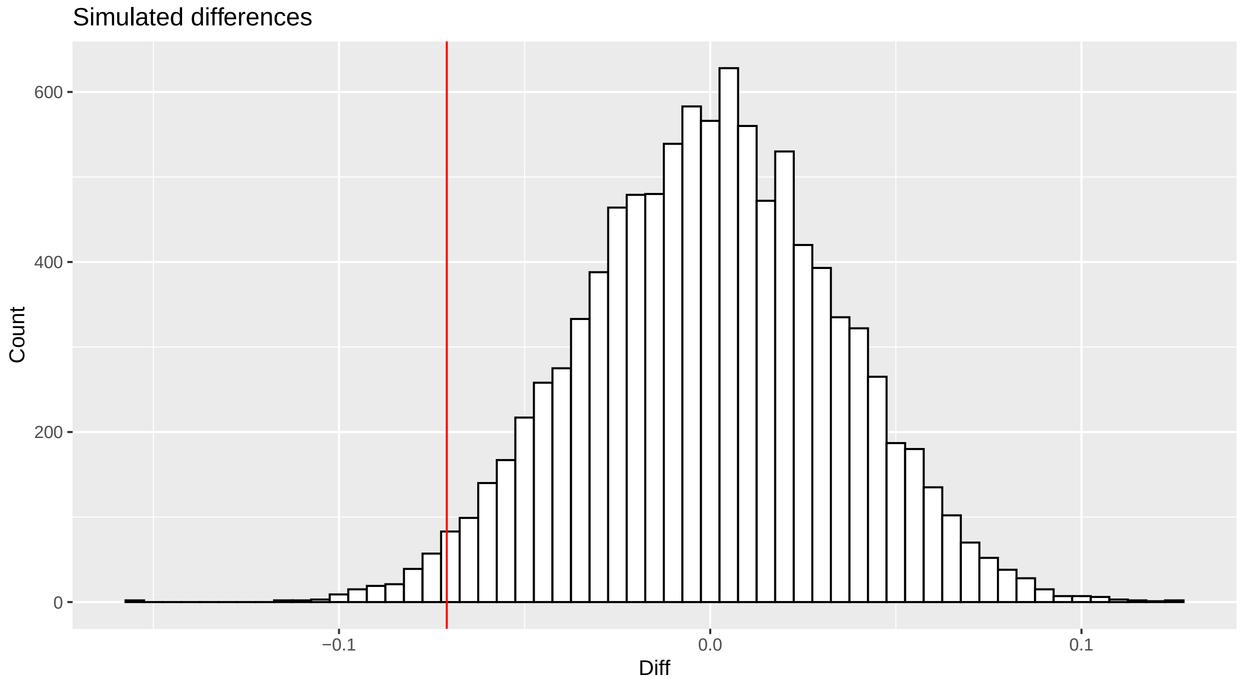
```
[1] -0.07096872
```

On average, non-isolated countries have HDI 0.0709 points lower than isolated countries.

Now we carry out a permutation test to check whether this difference is significant.

```
N <- 10000
diff <- numeric(N)

for (i in 1:N){
  samp <- sample(nrow(chisq_data), sum(No.border == 1))
  # obtain random sample of size equal to number of countries without a border in the data
  weightSamp <- mean(HDI[samp]) # mean for random sample
  weightOther <- mean(HDI[-samp]) # mean for complement
  diff[i] <- weightSamp - weightOther # calculate the difference
}
```

Let's plot a histogram of the results.



Since we did not try to establish the sign of the difference a priori, we should carry out a two-tailed test.

```
((sum(diff <= observed)+1)/(N+1))*2
```

```
[1] 0.03779622
```

This means that there it is quite unlikely (roughly 4.5% chance) that the observed difference happened by chance under the assumption that the null hypothesis of equal HDIs were true. We can thus confidently (at the 5% significance level) reject this null hypothesis and conclude that HDI levels for countries that do not have a border are significantly different from the HDI levels of countries that do have a border.

At this point, we can compare the Chi-squared test with the Permutation test.

The Chi-squared test is an "approximate test". As sample size n increases, the probability distribution of the test statistic $\sum_k \frac{(observed-expected)^2}{expected}$ converges to the Chi-squared distribution with $k-1$ degrees of freedom. However, when sample sizes are fairly small for some categories, as in our case, then the result might not be exact.

The Permutation test does not have this issue as it does not utilize any theoretical distributions in its functioning. With a sufficiently large number of simulations (resamplings), we can test the significance "exactly".

An exact test provides a significance test that keeps the Type I error rate of the test α at the desired significance level of the test. For example an exact test at significance level of $\alpha = 0.05$, when repeating the test over many samples where the null hypotheses is true, will reject at most 5% of the time. This is opposed to an approximate test in which the desired type I error rate is only approximately kept (i.e.: the test might reject more than 5% of the time), while this approximation may be made as close to α as desired by making the sample size big enough.

In summary, Permutation test (and other non-parametric tests including many using Monte Carlo techniques) are better when the sample size n is small.

5. The World Belligerence Index : Constructing a Global Index using PCA.

Abstract :

In this topic, we construct a “World Belligerence Index” using Principal Component Analysis. We have counts for a wide range of actions taken by over 200 world countries for a period ranging from 1990 to 2004. For each action type, we convert the counts to the percentage of total counts for that action accounted for by each country. We find that one of our Principal Components captures variance in the data that can be plausibly associated to a belligerent attitude in the international arena. The set of actions identified by the principal component is mostly composed of belligerent/aggressive actions. Using these, we construct an index of belligerence as an average of the percentages of total counts accounted for by each country for each action identified by our principal component. We then regress the Index scores on GDP per capita, HDI, Civil conflicts and a binary variable for presence of a border with another country. These data for these covariates were constructed in preparation to the topics previously covered in this project.

Datasets used : 10 Million International Dyadic Events (10MM_IDE).

I. Constructing the World Belligerence Index.

We narrow down the dataset to just 24 actions. We filtered the actions on the basis of relevance (to the purpose of the index) and removed actions with too low variance.

The overall activity level of countries in the international arena leads to a high correlation between action counts for all actions. (For example, the US, being a major global actor, has high counts on virtually all actions, good or bad). This biased the results of our initial attempt at carrying the PCA and motivates our decision to manipulate the data to use the proportion of total counts for each action instead.

```
    Countries AERI        BREL CLAS         COLL      COMP_1        DMOB       DWAR
1         AFG    0 0.006849315    0 0.0046189376 0.002544529 0.000000000 0.03389831
187       ALB    0 0.000000000    0 0.0000000000 0.013994911 0.000000000 0.00000000
373       ALG    0 0.006849315    0 0.0040415704 0.002544529 0.000000000 0.00000000
559       AND    0 0.000000000    0 0.0000000000 0.000000000 0.000000000 0.00000000
745       ANG    0 0.027397260    0 0.0005773672 0.000000000 0.006430868 0.01694915
931       ANT    0 0.000000000    0 0.0000000000 0.000000000 0.000000000 0.00000000
```

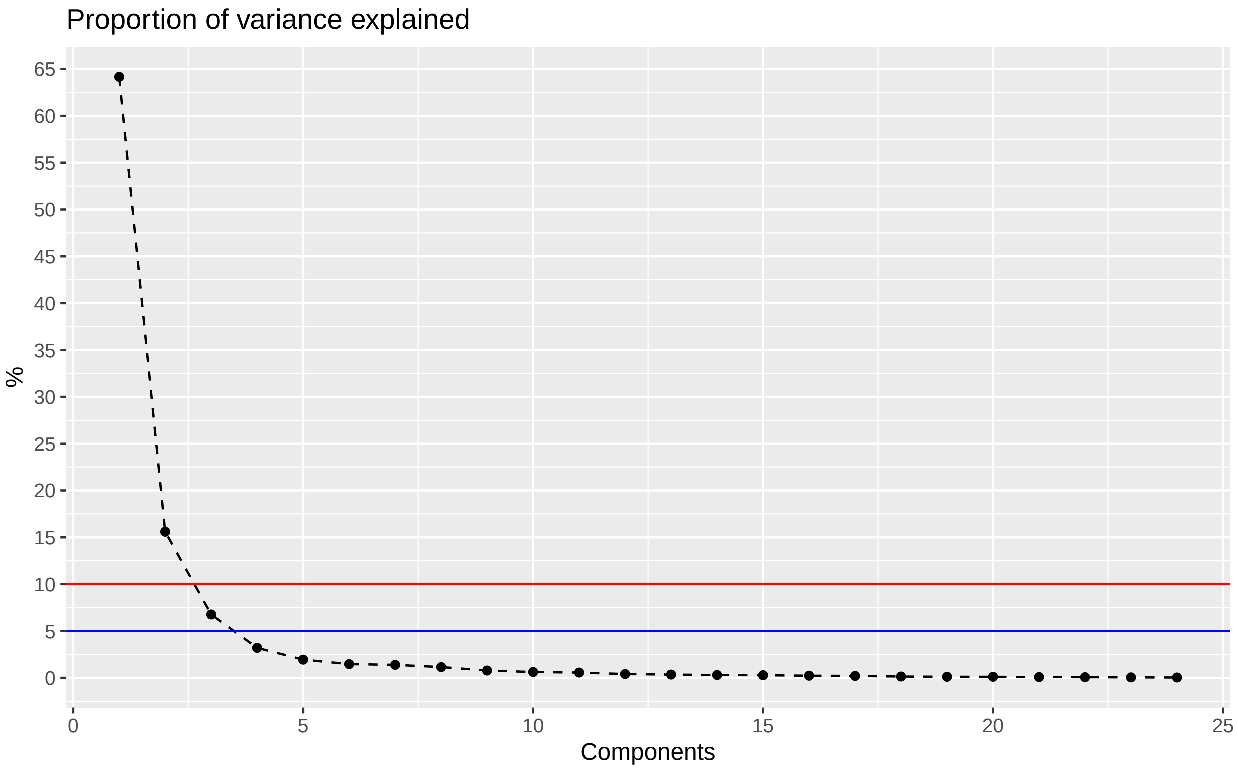
We use Eigenvalue decomposition to compute the Principal components. We create the symmetric covariance matrix and compute its Eigenvectors. These form the orthogonal eigenbasis of our covariance matrix by the Spectral Theorem.

```r
m_5<- nrow(PCA_data_5) ; m_5
A_5 <- PCA_data_5[,2:25] # Select relevant vectors
S_5 <- var(A_5) # Covariance symmetric matrix

#Now we have a symmetric matrix to which the spectral theorem applies.
Eig_5 <- eigen(S_5)
Eig.vals_5 <- Eig_5$values
P_5 <- Eig_5$vectors# This is the change of basis matrix, composed of the eigenvectors
```

These eigenvectors are our Principal Components. By the Spectral Theorem, we know they are orthogonal. We calculate how much of the total variance our components account for. We generate a Scree Plot and retain only the ones that explain more than 5% of total variance.

```r
## Lets calculate total variance and the proportions explained by each component
summ_var_5 <- 0
variances_5 <- numeric(24)
```

```
for (i in 1:24) {
  for (j in 1:24){
    if (i == j){
      summ_var_5 <- summ_var_5 + S_5[i,j]
      variances_5[j] <- S_5[i,j]
    }
  }
}
```



```
prop.var_5[1]
```

```
[1] 0.6415987
```

```
prop.var_5[2]
```

```
[1] 0.1559826
```

```
prop.var_5[3]
```

```
[1] 0.06764636
```

It seems like the first component, the one with the largest eigenvalue, accounts for over 64.1% of the total variance. The second and third account for 15% and 6.7% respectively.

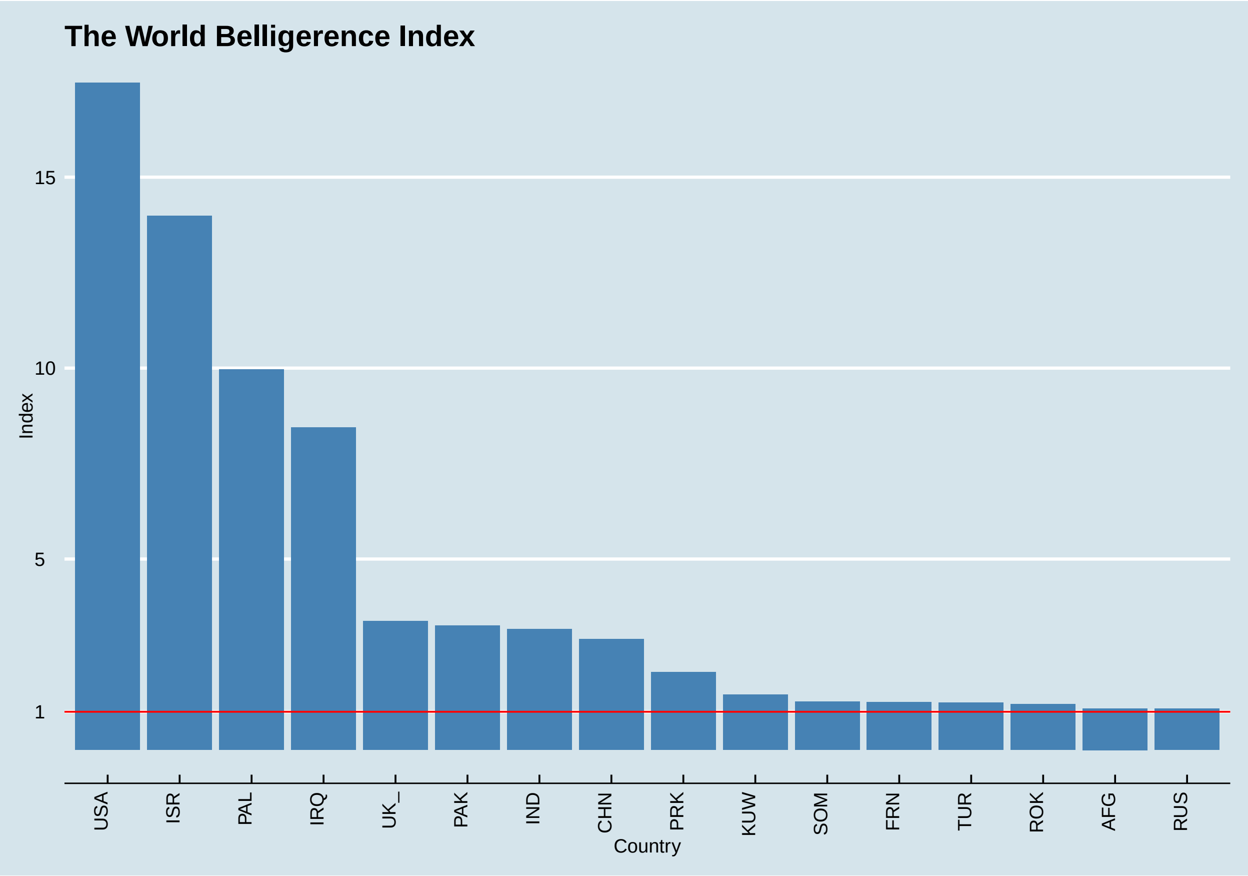
The first component PC1 has negative loadings for each action type and is likely accounting from something related to the overall activity level of countries in the international arena. PC2 and PC3 on the other hand do seem to capture something related to belligerance. We focus on PC2 since this accounts for over 15% of the total variation. On PC2 we see positive loadings on actions that we might normally associate to belligerant and/or aggresive nations. The actions with positive loadings are:

- RAID (Armed actions)

- POLPER (Political flight/arrests)
- AERI (Missile attacks)
- CLAS (Armed battle)
- PASS (All uses of non-armed physical force in assaults against people)
- EXIL (Expel)
- PEXE (Small arms attack)
- GRPG (Artillery attack)

Now we use the belligerent actions identified by PC2 to form our Index scores.
We take the arithmetic mean of the action percentages.

```
attach(PCA_data_5)
Belligerance.index <- as.data.frame(PCA_data_5$Countries)
Index <- (RAID+POLPER+AERI+CLAS+PASS+EXIL+PEXE+GRPG)/8
Belligerance.index$Index <- Index*100
Belligerance.index <- Belligerance.index[order(-Belligerance.index$Index),]
colnames(Belligerance.index) <- c("country", "Index")
```

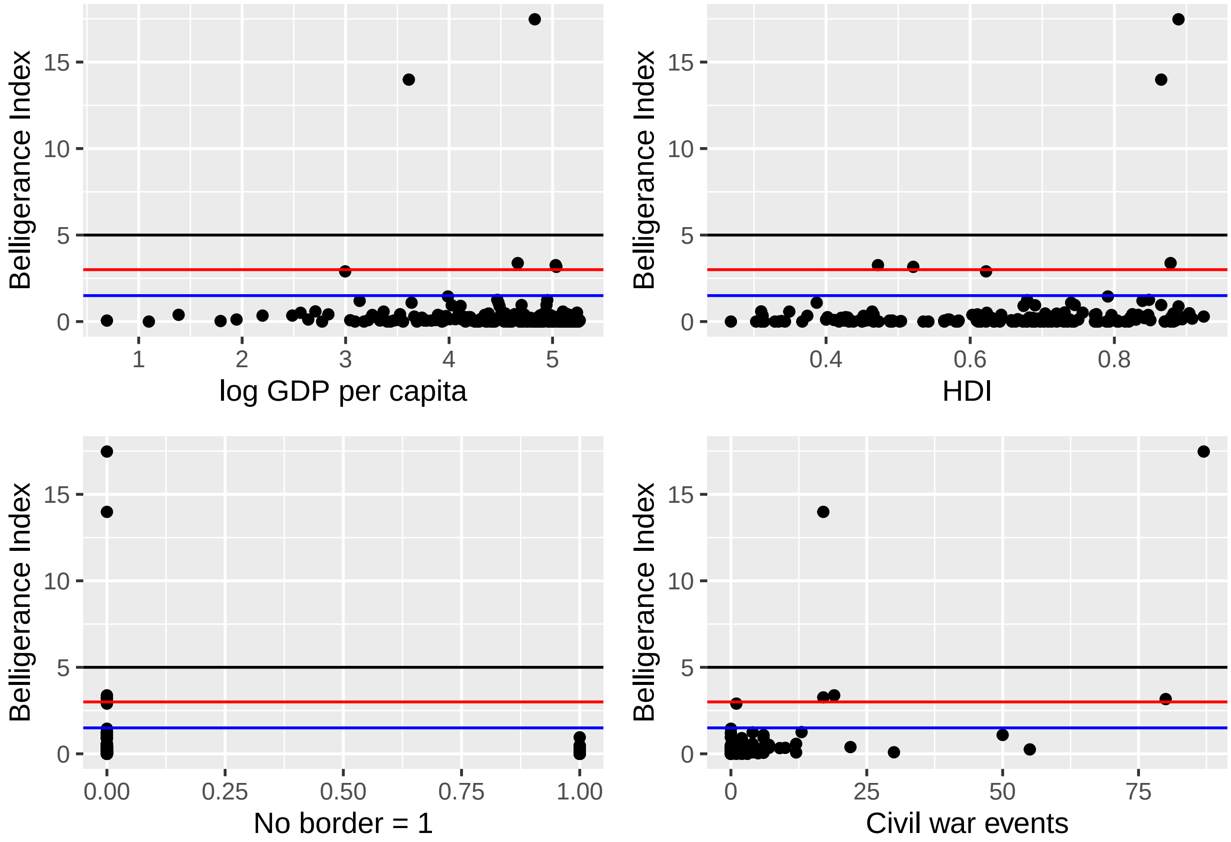


II. Linear Regression

In this section, we use Linear Regression to model how our Index relates to some of the variables we computed and used in all our topics upto now. We use data from the year 2003.

We construct the dataframe,

```
  country      Index      GDP CivilWars   HDI No.border
1     AFG 1.08873280 3.637586        50 0.387         0
2     ALB 0.09598991 3.496508         4 0.687         0
3     ALG 0.05605381 3.828641         0 0.676         0
4     AND 0.00000000 4.584967         0 0.820         0
5     ANG 0.09598991 5.257495         0 0.428         0
7     ARG 0.42628969 4.634729         1 0.775         0
```



We use the projection matrices approach to conduct Linear Regression.

$$\hat{\mathbf{y}} = A(A^T A)A^T \mathbf{y}$$

$$\hat{\mathbf{y}} = A\Phi$$

where, $\hat{\mathbf{y}}$ is the vector of projected y values, A is the matrix with $\mathbf{1}$ and the covariate vectors as its columns, and Φ is the vector of coefficients.

```
v1 <- c(rep(1, nrow(Reg_vars)))
A <- cbind(v1, Reg_vars$GDP, Reg_vars$HDI, Reg_vars$CivilWars, Reg_vars$No.border)
coeff <- solve(t(A)%*%A)%*%t(A)%*%Reg_vars$Index
```

We find that the built-in R function lm() returns the same results. We use it to find the t-statistics and significance of covariates in our model.

```r
reg_model <- lm(Reg_vars$Index ~ Reg_vars$GDP + Reg_vars$HDI
                + Reg_vars$CivilWars + Reg_vars$No.border)
```

```
===========================================================
                                   Dependent variable:
                               ----------------------------
                                   Belligerance Index
-----------------------------------------------------------
Log GDP per cap.                         -0.027
                                         (0.124)

Human Development Index (HDI)           2.353***
                                         (0.647)

Civil war                               0.098***
                                         (0.010)

Lacking border                           -0.323
                                         (0.291)

Constant                                 -1.233*
                                         (0.686)

-----------------------------------------------------------
Observations                               165
R2                                        0.421
Adjusted R2                               0.407
Residual Std. Error                 1.380 (df = 160)
F Statistic                      29.143*** (df = 4; 160)
===========================================================
Note:                          *p<0.1; **p<0.05; ***p<0.01
```

The results suggest that HDI and the occurrence of civil wars lead to increases in the level of belligerance shown by countries as measured by our index. However, given suspicion (motivated by the initial scatterplots of the dependent variable with each individual regressor) that a linear model might not be appropriate for this analysis, we advise not to read too much into these.

6. Fulfilled Project Requirements :

Required dataset standards :

- A dataframe. (Many)
- At least two categorical or logical columns. (Topic 4 : HDI, Border columns)
- At least two numeric columns. (Topic 3)
- At least 20 rows, preferably more, but real-world data may be limited. (10 million rows!, although we reduce it to include only inter-country events.)

Required graphical displays :

- A barplot. (Topic 5 : World Belligerent Index)
- A histogram. (Topic 1,2,3)
- A probability density graph overlaid on a histogram. (Topic 1,2,3)
- A contingency table. (Topic 4)

Required analysis :

- A permutation test. (Topic 4)
- A p-value or other statistic based on a distribution function. (Topic 1, Topic 4 : we calculate p-value using our chi-sq. statistic and based on the chi-squared distribution.)
- Analysis of a contingency table. (Topic 4)
- Comparison of analysis by classical methods (chi-square, CLT) and simulation methods. (Topic 4 : We discuss Chi-sq and Permutation tests.)

Required submission uploads :

- A .csv file with the dataset (Many)
- A long, well-commented script that loads the dataset, explores it, and does all the analysis.
- A shorter .Rmd with compiled .pdf or .html file that presents highlights in ten minutes.
- A one-page handout that explains the dataset and summarizes the analysis.

Additional points for creativity or complexity :

- A data set with lots of columns, allowing comparison of many different variables. (Many variables used in the entire project.)
- A data set that is so large that it can be used as a population from which samples are taken. (10 million observations in original dataset.)
- A graphical display that is different from those in the textbook or in the class scripts. (Many)
- Appropriate use of R functions for a probability distribution other than binomial, normal, or chi-square. (Topic 1 : Exponential, Weibull)
- Appropriate use of integration to calculate a significant result. (Topic 1 : Calculated Expectation.)
- A convincing demonstration of a relationship that might not have been statistically significant but that turns out to be so. (Topic 4 : Permutation test gives significant difference which logically shouldn't have existed.)
- A convincing demonstration of a relationship that might have been statistically significant but that turns out not to be so. (Topic 2,3 : Network analysis results have many insignificant results which might have been hypothesized to be true.)
- Professional-looking software engineering (e.g defining and using your own functions). (Many, examples are making own Weibull functions in Topic 1, vector generator functions in Topic 2.)
- Nicely labeled graphics using ggplot, with good use of color, line styles, etc., that tell a convincing story. (Many)
- An example where permutation tests or other computational techniques clearly work better than classical methods. (Topic 4 : We saw contradictory results for Chi-sq and Permutation test, discussed advantage of the latter in cases with small samples.)
- Appropriate use of novel statistics (Topic 1 : Likelihood Ratio test.)

- Use of linear regression. (Topic 5)
- Appropriate use of covariance or correlation. (Topic 5 : PCA uses a covariance matrix which is used for Eigenvalue decomposition.)
- A graphical display that is different from those in the class scripts. (Many)
- Team consists of exactly two members.
- A video of the short script is posted on YouTube and a link to it is left in your long script.

7. Conclusion

In this project, we utilized many statistical techniques as well as mathematical theory learnt in as well as outside class to analyze an interesting and rich dataset. We try to incorporate methods taught in class that require understanding of the underlying theory instead of using readymade functions.

In Topic 1, we explored how we could utilize the properties of the Exponential and Weibull probability distributions to find peer influence amongst countries with respect to certain actions. We found "armed assistance requests" to be peer influenced and demonstrated a counter-example of "ultimatums", which did not seem to be peer infuenced.

In Topic 2, we analyzed the network of inter-country threats and tried to find whether provision of economic aid had an impact on threat patterns. With some caveats, we found that providing economic aid to a country reduced the probability of receiving a threat from it, thereby implying that aid provision might be effective in "buying" softpower.

In Topic 3, we looked at the Political flight network to try to find some interesting factors that might influence the target locations of fleeing politicians. We found that politicians are more likely to flee from countries undergoing civil conflicts to similarly distressed countries. Moreover, we found that politicians are also more likely to flee to countries with the same major religion as their own country.

In Topic 4, we looked at the relationship between Geographical Isolation and Human Development levels. Even though there doesn't seem to be any logical justification to it, results from our permutation test found a significant difference in mean HDI between isolated and non-isolated countries. The Chi-squared test had contradictory results which was also discussed.

In Topic 5, we constructed the "World Belligerence Index" and used linear regression to find relationships between the Index scores and several other variables.

8. Acknowledgements

We would like to thank Paul Bamberg for making and teaching such an amazing course. It not only exposed us to many interesting ideas but also helped us dig into their intricate mathematical foundations.

We would like to thank Grant, Alek and all the TAs that made this course incredibly rewarding and accessible.

Last but not the least, We would like to thank Head TA Michael Liotti who worked incredibly hard to not only clear all our course-related doubts but also provide invaluable advice whenever we asked him. We will miss your "epic" Office Hours and Sections. A self-proclaimed "perfectionist", he has certainly perfected the art of TA'ing!

9. References

In addition to course-material, we used several R tutorials while making this project. The main dataset used in this project was Gary King's 10 Million International Dyadic Events.

- King, Gary; Lowe, Will, 2008, "10 Million International Dyadic Events", https://doi.org/10.7910/DVN/BTMQA0, Harvard Dataverse, V5, UNF:3:dSE0bsQK2o6xXlxeaDEhcg== [fileUNF]
- van der Pol, J. Introduction to Network Modeling Using Exponential Random Graph Models (ERGM): Theory and an Application Using R-Project. Comput Econ 54, 845–875 (2019). https://doi.org/10.1007/s10614-018-9853-2