Corrupted Data and the Illicit Arms Trade

Statistical Science Honors Thesis

Rose Graves

Thesis Advisor: Alexander Volfovsky

Duke University

Durham, NC

2020

Corrupted Data and the Illicit Arms Trade

Rose Graves

April 6th, 2020

Abstract

With the ever-increasing advancements in weapons technology, the illicit arms trade has steadily become a greater threat to international security. The small arms trade, consisting of portable weapons and their parts, is not only a profitable good, but also a method of gaining power through violent and threatening means. Being able to identify when and what countries are engaging in illicit arms trade is essential in order to make informed policy decisions. The driving question behind this project is: how do we recognize corrupted network data and how does corrupted network data impact our statistical analyses? The arms trade takes the form of network data consisting of actors (nodes) and the relationship between them (edges). This analysis of methods initially looks at simulated data. We show that if data is sampled from a pre-specified model then increasing the amount of corrupt data present impacts posterior statistics such as the intercept and row and column coefficients, as well as posterior predictive descriptive statistics such as degree distributions, triangle counts, betweenness, and Eigen vector centrality. This analysis demonstrates if data is corrupt, then by replacing the corrupted values with NAs these missing values will be imputed from the true pre-specified model and thus will not impact inference. These methods are then applied to actual small arms trade data, to see what nations may be engaging in illicit arms trading.

Thanks

I want to initially thank my advisor Alex Volfovsky for assisting me at every step with this project. I'm so appreciative of your patience and willingness to help me undertake this research for the past year. I want to thank Heather Mathews for the taking the time to answer my abundance of questions throughout the past year. I truly don't think I could have produced this without your help. Thanks to Joan Combs-Durso who has been a wonderful source of support and my personal cheerleader. Lastly, I want to say thanks to all the friends and family who have proof-read my work, brought me coffee, wrote me encouraging notes, or were just supportive throughout this process. Thank you, Emma Shaffer, Pablo Piedra, Arthi Kozhumam, Vicki Zhang, and my mom.

Chapter 1

Introduction

With the ever-increasing advancements in weapons technology, the illicit arms trade has steadily become a greater threat to international security. The small arms trade, consisting of portable weapons and their parts, is not only a profitable good but also a method of gaining power through violent and threatening means. The United Nations Office on Drugs and Crime (UNODC) notes multiple ways in which nations and criminal groups end up with illicit arms, such as fake export arrangements and incorrect recipients. The UNODC notes in Legal Firearms Market that firearms transfers must be documented alongside specifications of what is being shipped and the target country (UNODC, 2019). However, these arms transfers are sometimes labeled instead as machine parts and lead to the underreporting the true number of arms that are being transferred, or they are sent to a different nation than the one reported (UNODC, 2019). The United Nations estimates that the value of the illegal small arms trade may run into the billions, but a majority of the current literature on the small arms trade has failed to account for these unreported arms in their data (UNODA, n.d). The Illicit small arms trade encompasses criminal groups, non-state actors, and nations; however, this paper will be focusing on illicit small arms trades between nations. It is important to note the emphasis on small arms rather than all types of arms trading. The data on large arms trades are considered relatively accurate and complete, given the difficulty of covertly transporting weapons of such magnitude. Illicitly trading a box of handguns presents fewer transportation challenges than illicitly trading a tank, missile, or submarine.

During the conflict in former Yugoslavia in 1992, there was an arms-embargo placed on all the Yugoslav states by the United Nations. Bosnia, Croatia, and Kosovo had to engage with other nations and non-state actors in order to import weapons. Their main suppliers were Iran and Chile who violated the embargo and sent the weapons disguised as other goods or specified the incorrect recipients. A shipment of arms from Chile to Yugoslav states was disguised as humanitarian aid resources for Sri Lanka (Arsovska, 2014). This is an example where nations cheated on their obligation to accurately and honestly report their arms trades. It demonstrates the ways in which nations underreporting their arms trades can impact entire populations and hinder our ability to understand the consequences and causes of the arms trade.

Despite the large impacts illicit arms trades have on nations and individuals, the political science literature addressing the illegal arms trade or attempting to predict or correct for illegal arms trading is scarce. The research within the arms literature has been narrowly focused on particular regions or time periods out of concern for inaccurate data. This has greatly impacted both the quantity and quality of literature around this integral aspect of international relations. Having accurate information on arms trading is essential in order for policy makers to make well informed decisions and assess safety concerns in nations where governments are trading illegally. These concerns and questions have guided our research questions: how do we recognize corrupted network data and how does corrupted network data impact our statistical analyses?

Corrupted Data and the Arms Trade

Corrupted network data, specifically in the context of illegal arms trading, typically presents itself in two forms. The first is the failure to report any weapons transfers. Data points in this context are simply represented as "NA". The second is reporting no weapons transfers or a fewer number of transfers than actually took place. The presence of NAs in an arms trading network typically arises where there is a lack of infrastructure that allows such things to be reported. Currently there are multiple methods that deal with missing at random data that have shown to be very effective. Multiple Imputation by Chained Equations (MICE), has risen as one of the most popular methods for addressing missing data. MICE assumes that the values in the data are missing at random, or that the missing value only depends on observed values and is not related to other unobserved values. A conditional distribution is placed upon the variables in the data, and through Gibbs sampling datasets are created from these distributions (Buuren, 1999). In our analysis we are not implementing MICE, but rather setting up a full generative model and sampling from it; these are parallel problems. However, the main issue arises when countries are underreporting their arms trading, resulting in corrupted data.

Nations who are under an arms embargo or who may not want to report all of their arms trading due to possible conflicts of interest, may under report the number of weapons they bought or sold. Thus, while there are clear violations of arms embargos, there can be even greater complications when weapons are exported to states that are geographically close to an embargoed state, to then be transported to that embargoed country (Rothe & Collins, 2011). Thus, states engaging in illicit arms trading will under report or not report at all the number of weapons they are truly trading. Arms databases, in particular the Norwegian Initiative for Small Arms Trade (NISAT), while providing very rich sources of information, are not capturing all arms trades between nations. The NISAT database contains information on arms transfers between states and regions, the types of weapons traded, and the source of the information from 1962 to 2015.

Arms trade takes the form of network data. Network data presents itself in the form of a matrix, sometimes called a socio-matrix, where there are nations (nodes) that are connected to each other through arms trading (edges). While previous work has been done to address data with an abundance of 0s, such as contaminations models including zero-inflated Poisson, there has been a gap in the literature when it comes to data in network format. Thus, this analysis aims to demonstrate how this corrupted network data impacts our statistical analyses and how we can recognize data that may be underreported.

Chapter 2

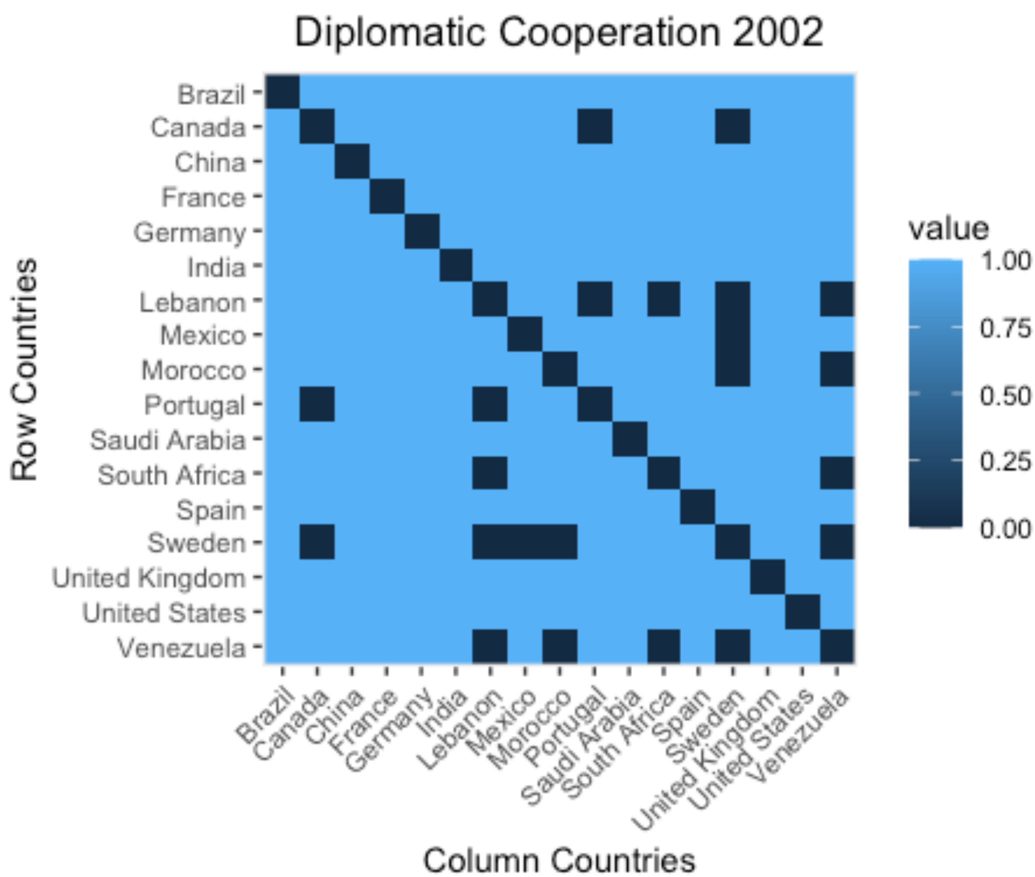
Literature Review

Networks Literature Review

Since arms trading data takes the form of network data, it's important to have a comprehensive overview of the literature and format of network data in order to understand how this data may be compromised. The notation and assumptions made in this section will stand for the rest of the paper. Social network models take the form of various actors (nodes) connected through a specific relationship (edges). For example, let us denote Y as an $n_x n$ matrix where each entry in the matrix represents an interaction between nations. If we examined whether two nations participated in diplomatic cooperation with each other, then our Y, also known as a socio-matrix, might look something like:

	Brazil	Canada	China	France	Germany	India	Lebanon	Mexico
Brazil	0	1	1	1	1	1	1	1
Canada	1	0	1	1	1	1	1	1
China	1	1	0	1	1	1	1	1
France	1	1	1	0	1	1	1	1
Germany	1	1	1	1	0	1	1	1
India	1	1	1	1	1	0	1	1
Lebanon	1	1	1	1	1	1	0	1
Mexico	1	1	1	1	1	1	1	0

This socio-matrix above displays a sample of nations and whether they engaged in diplomatic cooperation, 1, or not, 0, in 2002. Notice that along the diagonal all the values are zero, this is because nations cannot engage in diplomatic cooperation with themselves. Below is another representation of a socio-matrix, where light blue represents that the nations engaged in diplomatic cooperation and dark blue means that they did not.

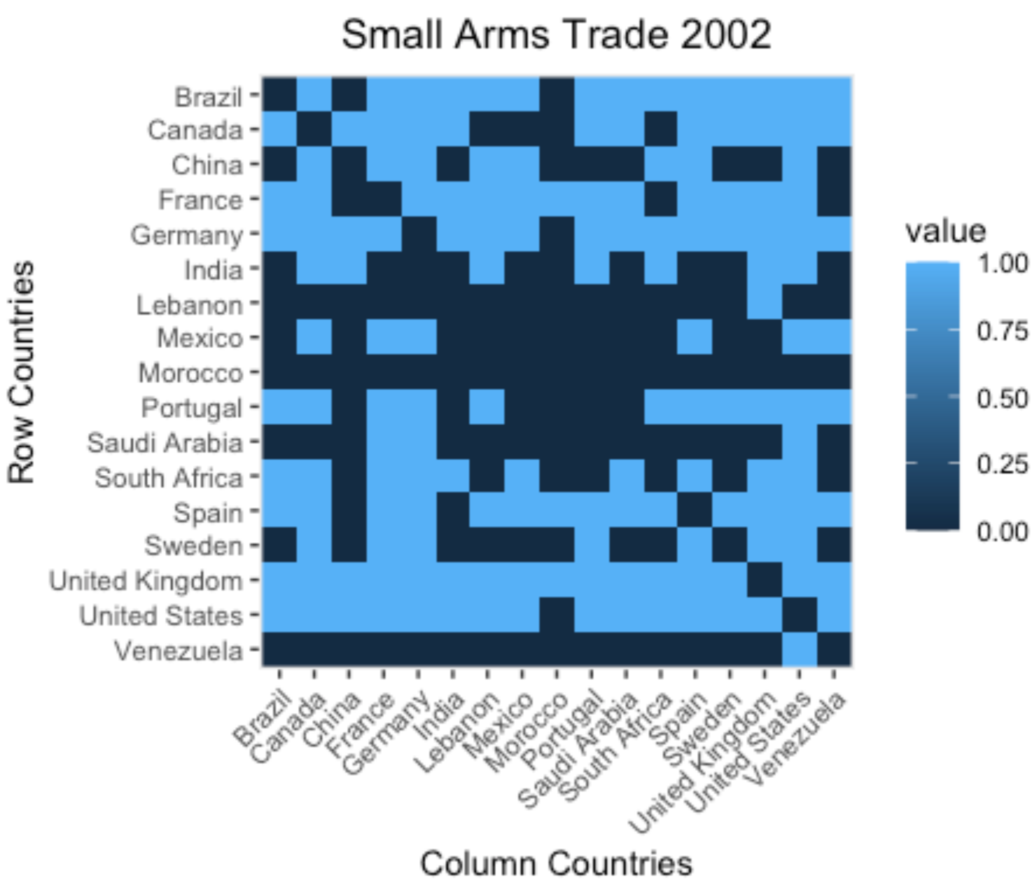


Here an entry of 1 represents that there is a relationship present. In this case, that the two nations engaged in diplomatic cooperation with each other. An entry of 0 represents a lack of a relationship. This is an example of a symmetric matrix, where $Y_{Germany,France} = Y_{France,Germany}$. This implies that if Germany participated in diplomatic cooperation with France, then France participated in diplomatic cooperation with Germany, and thus we get a symmetric matrix, where $Y^T = Y$.

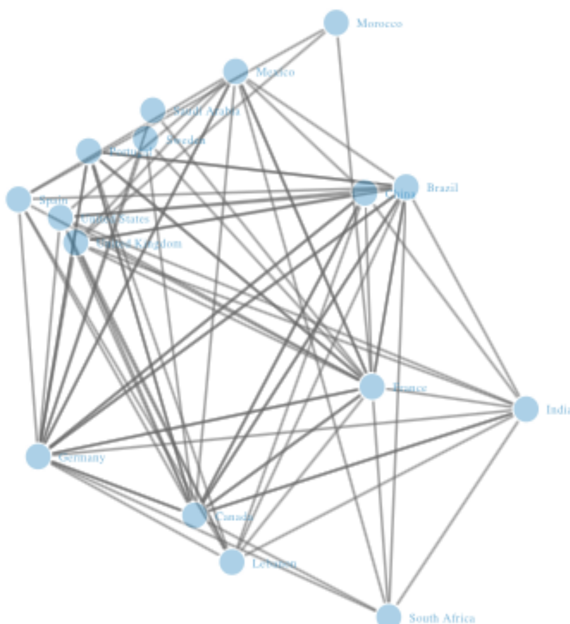
In the case of most trade data, the relations are not reciprocated, hence the socio-matrices are not symmetric. For example if Germany sent arms to France, $Y_{Germany,France}$ = 1, then it is not necessarily true that France sent weapons to Germany, and $Y_{France,Germany}$ could take the value of either 1 or 0. Networks of this type are labeled as directed, and in this case our socio-matrix may take the form of:

	Brazil	Canada	China	France	Germany	India	Lebanon	Mexico
Brazil	0	1	0	1	1	1	1	1
Canada	1	0	1	1	1	1	0	0
China	0	1	0	1	1	0	1	1
France	1	1	0	0	1	1	1	1
Germany	1	1	1	1	0	1	1	1
India	0	1	1	0	0	0	1	0
Lebanon	0	0	0	0	0	0	0	0
Mexico	0	1	0	1	1	0	0	0

This socio-matrix above contains a sample of nations and whether or not they sold arms, 1, or not, 0, and whether they bought arms, 1, or not,0, in 2002. Below is another visual representation of the socio-matrix.



These relationships, $y_{i,j}$, denote the relationship from actor i to j and can be visualized in the form of a network graph where each node represents an actor, and each edge represents a relationship (directed or non-directed, weighted or non-weighted) between the actors. Below is an example of a network graph, where the nodes are a sample of nations, and the edges represent that an arms trade occurred in the year 2002. This contains the same information as the previous socio-matrix but is a different visual representation.



In the network graph above, the nodes represent the nations, and the edge between nodes represents that arms trading occurred between them.

When outcomes are continuous, relationships and data of this type can be modeled with a similar approach to linear regression:

$$y_{i,j} = \beta^T x_{i,j} + \epsilon_{i,j}$$

Here, x represents an array of covariates, and ϵ is the error, which is potentially non-independent. The co-variates can be on a nodal or dyadic level. For example, on the nodal level this can be information about actors such as GDP for a country. Dyadic covariates, however, take the form of a socio-matrix.

In the case of binary data, we represent this type of data in the following form.

$$z_{i,j} = \beta^T x_{i,j} + \epsilon_{i,j}$$

Where z represents some continuous latent variable, and implementation of an indicator function will determine which entries of z will be represented as 1s and which ones will be 0s:

$$y_{i,j} = 1(z_{i,j} > 0)$$

Thus, in this case, the entries of $z_{i,j}$ that are greater than 0, will become 1.

However, there were growing concerns over the applicability of these linear models, given that they fail to take into account dependencies present in dyadic data. Since the ϵ is not necessarily independent, we can't simply vectorize the data as there is structure and dependency we want to account for. If data is sufficiently structured, then it can give us insight into how one can better estimate parameters.

If we consider the heterogeneity present in the rows and columns of the dyadic variables, then if the rows represent the countries who are exporting goods and the columns represent the importers, we would expect there to be difference among how much each nation sells and how much nations are purchasing. These are considered second order dependencies, which capture reciprocity and within-actor correlation (Hoff & Ward, 2004). These row, column, and dyadic effects take the following form.

$$\epsilon_{i,j} = a_i + b_j + \gamma$$

Where,

$$\begin{pmatrix} a \\ b \end{pmatrix} \sim Normal\begin{pmatrix} 0 & , & \sigma_a^2 & \sigma_{ab}^2 \\ 0 & & \sigma_{ba}^2 & \sigma_b^2 \end{pmatrix}$$

$$\begin{pmatrix} \gamma_{i,j} \\ \gamma_{j,i} \end{pmatrix} \sim Normal\begin{pmatrix} 0 & , & \sigma_\gamma^2 & \rho\sigma_\gamma^2 \\ 0 & & \rho\sigma_\gamma^2 & \sigma_\gamma^2 \end{pmatrix}$$

Hoff (2002) took a latent space approach to social network analysis expanding upon the Stochastic Block Models (SBM) method of clustering. Hoff discusses the concept of a social space consisting of latent characteristics representing transitive tendencies within the network. Incorporating a third order dependence allows for the ability to differentiate between ingroup and outgroup ties, by using transitive tendencies to represent latent characteristics of nodes and edges (Hoff, 2002). An example of transivity in context of arms trading can be shown if we consider three nations, A, B and C. Provided that A and B are trading together, then if B and C are trading together this impacts the probability that A and C are trading together.

Within international relations there are dependency structures that should be accounted for. One approach considers multiplicative effects which uses inner products of latent unobserved characteristic variables, and hence is able to incorporate third-order dependencies into the model (Hoff & Ward, 2004). These third order dependencies let us observe clustering properties such as transitivity and balance. This social relations regression model takes the form of:

$$\epsilon_{i,j} = a_i + b_j + \gamma_{i,j} + z_i^T z_j$$

where $\gamma_{i,j}$ is the dyadic error and remains independent of the bilinear and random effects, and $z_i^T z_j$ are the third-order dependencies.

The goal is to explore if we can detect misreported information, and in the cases where data had been misreported, understand how we can leverage the structure to better estimate parameters; therefore, it is important to consider various summary statistics and parameters that we think may be altered by corrupt data and how they may be impacted by corrupted data.

Degree Distribution

The degree of a node refers to the number of relationships that node has to other nodes. The degree distribution is the distribution of the degrees of all the nodes in the network. When considering directed networks, we can differentiate between in and out degrees. In degrees are the number of relations that point towards that node, while out-degree is the number of relations that originate at that node and point outward. This introduces the concepts of sociability and popularity. Considering the arms trade, a country would be considered sociable if it exported weapons to a large number of nations, and thus have a high out-degree. A country would be considered popular if it imported weapons from a large number of nations, and therefore possess a high in-degree.

We predict when nations underreport the number of weapons transferred then nations will appear to have less connections to other nations, and thus the number of degrees within that distribution will decrease, making the network seem scarcer than it truly is.

Triangle Count

A triangle, in the context of networks, refers to a cluster of three nodes that all have a relationship to one another. Counting triangles is a clustering algorithm and refers to the number of triangles that a node is a part of. Underreporting of arms may cause the network to appear less connected and change the structures of clusters in the network.

Betweenness

Betweenness measures how often a node lies the shortest on the path between other nodes. Nodes with high betweenness may have higher control over the network as more information, or goods, passes through it. If nations are underreporting their arms transfers then we would expect their betweenness to decrease, and thus it would seem as though they have less control over the network than they truly do.

Eigen Vector Centrality

This type of centrality is based upon the idea that important nodes are connected to other important nodes. Hence, if a nation is underreporting their arms, then they might seem less influential or less central to the network than they actually are.

Arms Trade Literature Review

In the previous section we saw that arms trading data has a directed and weighted structure, and thus an effective method to capture this information is through social network models. There has been very limited work within the political science literature to understand when and what nations have been trading illegally and even less work within the statistical science literature to address these issues.

Much of the previous work that has been done in examining the illicit arms trade has been through explanatory data analysis, case studies, and policy recommendations. The nature of the illicit arms trade and the reasons why states engage in the illicit arms trade has altered over time. If we consider the period of the Cold War, illicit arms trading was motivated by ideological and policy fears (Phythian, 2000). It was used as a tool to undermine nations with opposing ideologies. However, after the fall of the Soviet Union, the arms trade became relatively depoliticized and became a source of profit for many nations, nonstate actors, and regions fueling regional, ethnic, and religious conflicts (Phythian, 2000). Arsovka (2008) further describes the various methods and motivations for engaging in illicit arms trading, specifically within the Balkan region, such as profit for criminal groups, arming of individuals for terrorist or criminal related activities, and state engaged illicit arms trading. They note the importance of considering the multiple motivations for nations engaging in the illicit arms trade such as economic, social, cultural, and political reasons. This trend of changing relations in the small arms trade following the end of major military interventions and de-escalation of conflicts is also observed in the Middle East upon the intervention of the United States and throughout the gradual removal of the United States presence in the region (Bichler & Malm, 2013). Changes in the out-degree distributions and transitivity in the Middle East during the period of 2003-2008 imply that fractions are forming with a few key actors.

Other work has been done by the United Nations and other governmental research agencies who have made policy recommendations as to how to address illicit arms trading. These include recommending states' need to adhere to embargos and punish those who don't and strengthen monitoring of legal sales, so they don't get diverted to illegal or black markets (Stohl, 2004). There have also been acts by the United Nations and member nations to increase transparency of arms trading. The Arms Trade Treaty (ATT), which was put into place in 2014, has the goal of regulating arms transfers from small conventional weapons to large combat machinery. The United Nations acknowledges that the possession of weapons within a country's security force is essential in maintaining the order and protecting the rights of its citizens. However, it is of the upmost importance that nations ensure that the transfer and storage of the weapons are done so with the highest security (UN Office for Disarmament Affairs). The ATT attempts to address these concerns by requiring countries to implement a variety of measures to increase the transparency of arms transfers. Nations, under the ATT, are expected to establish a system to control exports, regulate imports and trans-shipments, and regulate brokering (United Nations

Office for Disarmament Affairs). While this body of work greatly increases our understanding in the motivations of actors to engage in illicit arms trading and the conditions that promote illicit arms trading, they fail to show us how we can tell when illicit trading is occurring and how this impacts our inference. It's difficult to make policy recommendations when it's unclear where illicit arms trading is occurring, when it is occurring or to what extent it is occurring. How can we punish states who are breaking arms embargos if we can't identify those who are breaking them? In order to efficiently allocate resources to addressing and stopping illicit arms trade we need to first understand the structure of the market.

Economic models have been used to see how fluctuations in stock prices may indicate possible illicit arms trading. By analyzing nations that were under arms embargos, it may provide insight as to who may have violated these sanctions by purchasing or selling arms. The assumption is that if a company is not partaking in illicit arms trading, then an increase in hostilities should either not affect the stock price or cause the stock price to decrease as it may lead to a prolonged embargo. However, if the company is engaging in the illegal arms trade then an increase in hostilities will increase the demand for weapons and in return may increase their stock price. Companies located in areas of high corruption and low transparency see an increase in their stock prices after an event that increases conflict intensity (DellaVigna, Stefano, & Ferrara, 2010). Looking at arms embargos provides a clearer method of hypothesizing what nations may be buying illicit arms. Arms embargos have the goal of decreasing the military activities of nations by cutting off a supply of power. Though reported arms trades typically decrease after the implementation of an embargo, illicit trading may occur and undermine the initial goals of the embargo (Brzoska & Lopez, 2009).

Arms embargos are typically placed upon nations where armed conflict or crimes against humanity are occurring or have occurred. These restrictions can create a large demand for arms, and for bordering nations this can provide an opportunity to engage in an illicit market with greater ease than nations who are geographically more distant. Radford (2013) finds that if nations that have at least one neighboring country with an embargo this leads to a 38% predicted increase in the number of small arms and light weapons imports. This provides evidence that neighboring nations may be importing a greater number of weapons to divert these weapons for resale in the embargoed nations.

As described above, international relations have a multitude of dependency structures and these are seen within the arms trade as well. Arms trade can capture latent features of relationships. Arms trading dyads have a very specific dynamic that goes beyond just similar membership in international organizations, political beliefs, and the trading of public goods. It represents a type of constrained hierarchy, where nations are contained within a cluster and have a place on the hierarchy of that cluster (Beardsley, in review). Stronger states and weaker states will engage in hierarchical relationships with each other, allowing stronger states to more easily control outcomes while weaker states may view the larger state as a form of protection and a measure of deterrence (Beardsley, in review). Thus, these dependencies indicate the importance of incorporating latent variables in our models when simulating datasets.

Chapter 3

Methods

Dealing with Corrupted Data

To reintroduce the question, when countries underreport their true arms trading, how does this impact our analyses and how can we spot nations that are illicitly trading? If the arms trade takes the form of a generative model, then we can leverage the structure of this model to notice when corrupt data may be present. If we sample data from a generative model, then replacing a proportion of the entries with NAs should not change the inference but may increase the uncertainty around the inference. Thus, if there are nations we believe to be incorrectly reporting their arms, then by changing their data into NAs those values will be imputed from the generative model and thus our inference should not drastically alter. However, when there are a greater proportion of 0s present in the data than would typically be sampled from the generative model, representing underreporting, then we expect there to be some differences in the inference between this corrupted data and the observed and data replaced with NAs.

It's important to note that when we replace the trade of the corrupt nations with NAs, we are making the assumption that the data is missing at random. This implies that there is a relationship between the observed data and the propensity of the missing data, however the actual values of the missing data are not related to them being missing. This may not be a meaningful assumption, and the data may truly be missing not a random. This implies that the propoensitity of the missing data is related to the missing values. For example, this may be the case if countries that are trading illicitly are importing a higher proportion of weapons or exporting a higher proportion of weapons to each other than the rest of the population. If this is the case, further research will need to be done on data that is not missing at random to address when this issue may arise. For the moment, we are also making the assumption that countries who are trading illicitly are only trading illicitly with other countries who are trading illicitly. Thus if Saudi Arabia and Pakistan are both trading illicitly then the values of $Y_{SaudiArabia,Pakistan}$ and $Y_{Pakistan,SaudiArabi}$ may be considered corrupted, but if Canada is considered to be trading legally then the values of $Y_{Canada,Pakistan}$ and $Y_{Pakistan,Canada}$ are considered true and coming from the overarching generative model. To address the main questions at hand, we will consider simulated data and then apply our methods to actual arms trading data.

Generative Model and Simulations

To first address the question of how corrupted data impacts our analyses, we will be creating a generative model and simulating data from it. This generative model will be considered the model from which true arms trade data is sampled from. We are going to be considering three main types of data in the analyses that follow. The fully observed data coming from the generative model, data with a proportion of the values that have been replaced with NAs, and data where a proportion of the values that have been replaced with 0s, signifying underreporting. As stated above we expect the observed data and data with a proportion of the values replaced with NAs to perform similarly when conducting inference. We initially created a generative model and sampled 10, 300_x300 , Y matrices. The generative model takes the form:

$$z_{i,j} = \beta^T x_{i,j} + a_i + b_j + \gamma_{i,j}$$

$$y_{i,j} = 1(z_{i,j} > 0)$$

The model was set up so that 15 percent of the entries in the matrix are 1s and 85 percent are 0s. After calculating the Z matrix, we then calculated the 85% quantile and subtracted that value from all elements in the matrix in order to obtain 15% of the values to be greater than 0, and 85% of the values to be less than 0. A nodal covariate was sampled from a $Normal(0,1)$, this nodal covariate was used as both a column and a row covariate. The coefficients for each of these covariates is set to 1. The additive row and column effects are from a multivariate normal.

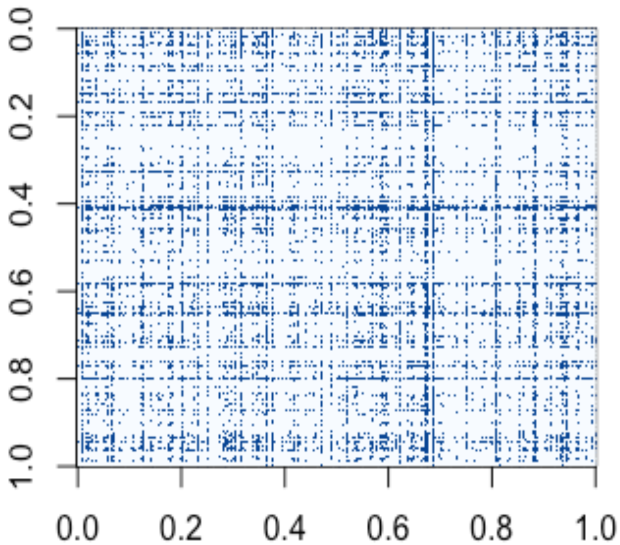
$$\binom{\alpha}{\beta} \sim Normal \begin{pmatrix} 0 & , & 1 & .5 \\ 0 & & .5 & 1 \end{pmatrix}$$

The errors are also pulled from a multivariate normal.

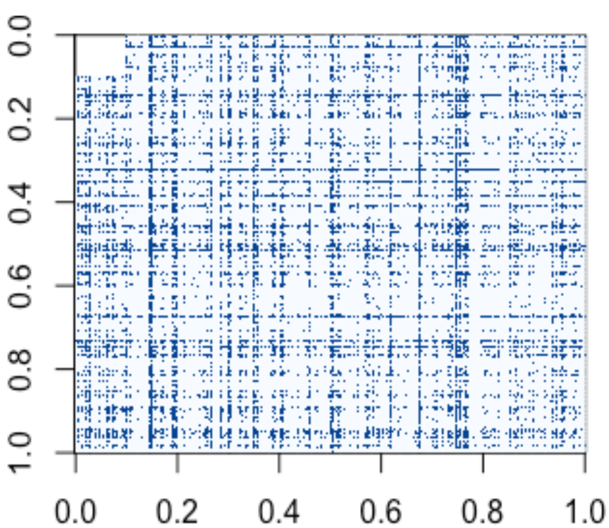
$$\binom{\gamma_i}{\gamma_j} \sim Normal \begin{pmatrix} 0 & , & 1 & .9 \\ 0 & & .9 & 1 \end{pmatrix}$$

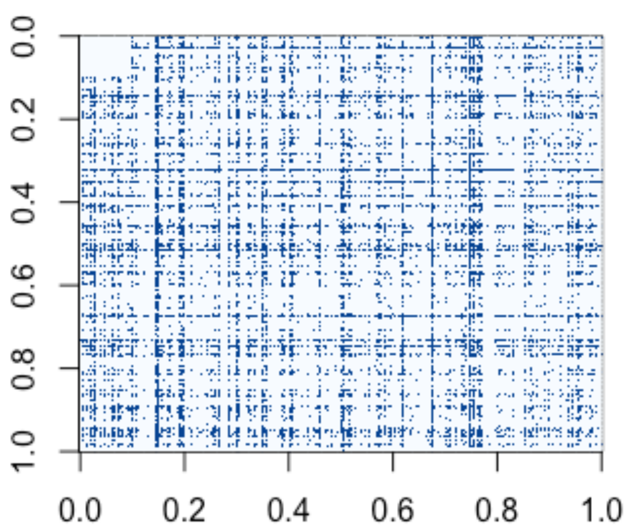
Multiple copies of each of these 10 datasets were made, and we replaced a percentage of the values in the data with either NAs or 0s. We replaced the data at varying percentages of 10 percent, 20 percent and 30 percent. By replacing a proportion of the values with 0s we hope to replicate how data is under reported. We are interested in how increasing levels of missingness and corruption impact certain posterior statistics and distributions. Below is an example of one of the simulated dataset and the modifications that were made.

The matrix below is one of the ten sampled matrixes without any modifications. This data is sampled from the generative model described above. The dark blue squares indicate that there is a relationship between two nodes and then the light blue squares represent the lack of a relationship between two nodes.

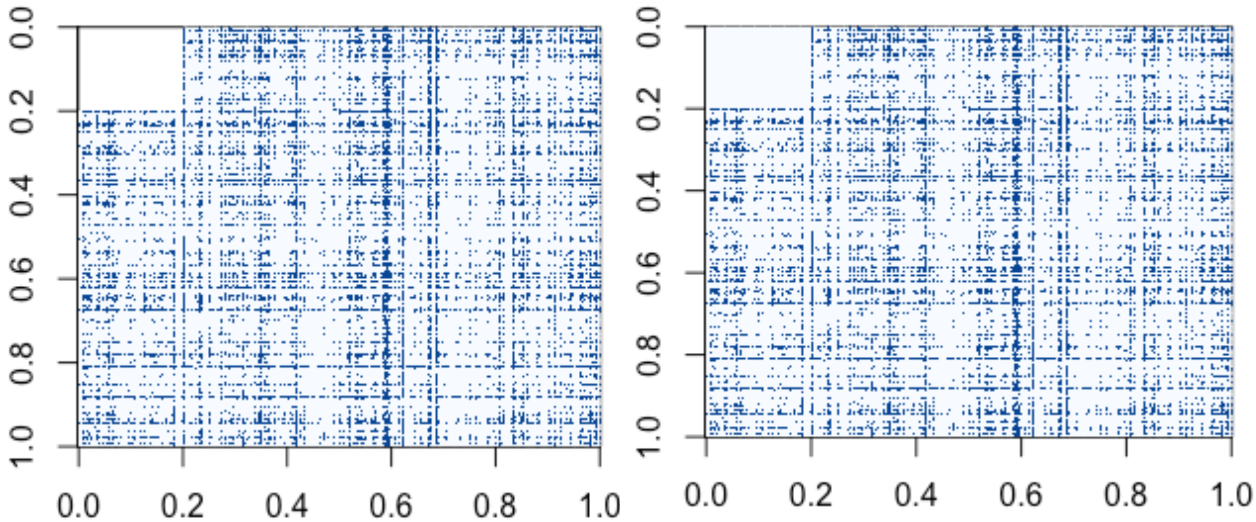


The two matrices below have been modified from the original sampled matrix. The matrix on the left has 10% of the entries replaced with NAs, this is indicated by the white square in the left corner. The matrix on the right has 10% of the values replaced with 0s.

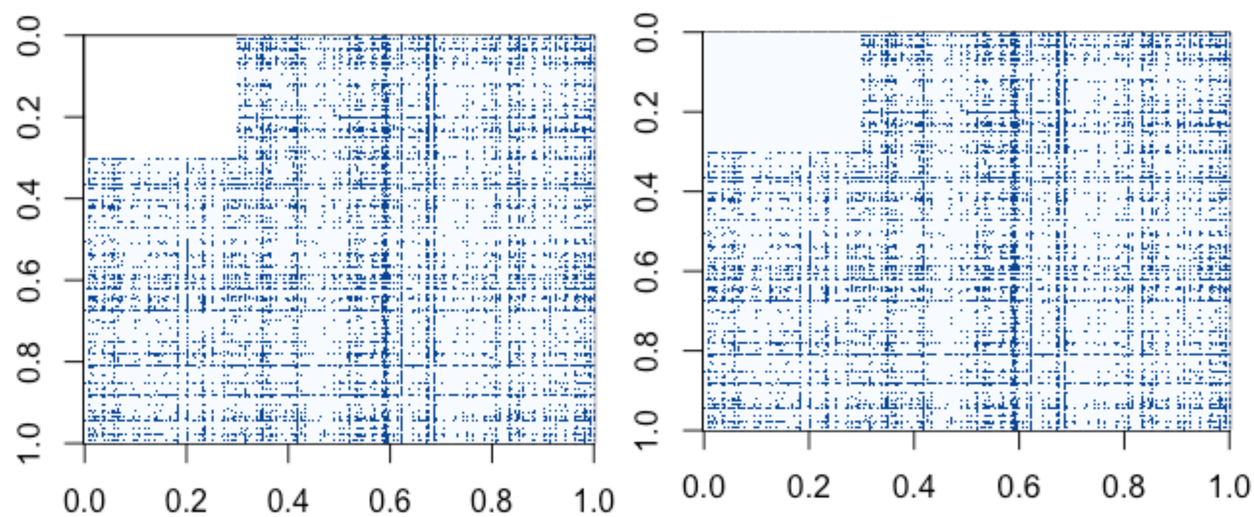




These two matrices have the same modifications as above, but the proportion of NAs and extra 0s in the matrix is 20%.



These two matrices have the same modifications as above, but the proportion of NAs and extra 0s in the matrix is 30%.



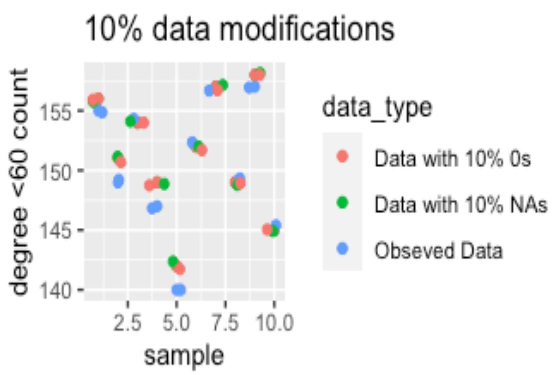
Exploratory Data Analysis

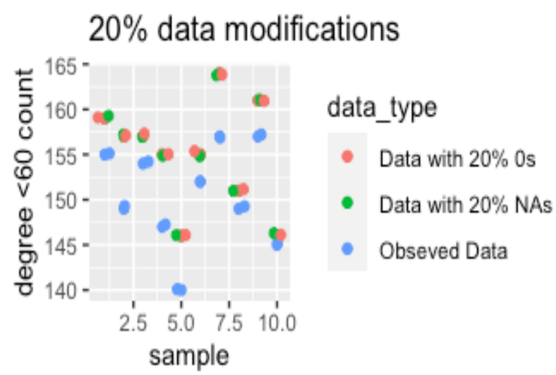
Before running social network models on the observed and corrupted data, a brief exploratory data analysis was completed. The graphs may give us an indication of what we expect the posterior distributions to give us.

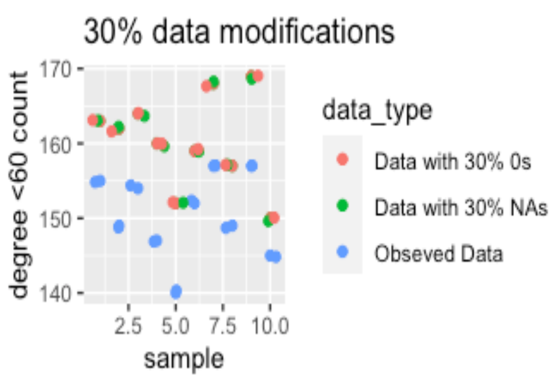
Degrees

All-degree

The graph below displays the number of actors with total degrees less than 60, across 10%, 20% and 30% replacements with NAs or 0s. As we increase the proportion of corrupted data there is a clear gradual widening in the difference between the values of the number of nodes with a degree less than 60. In the first graph all the three points for each sample are very similar but as we increase the amount of corrupted data the gap widens.

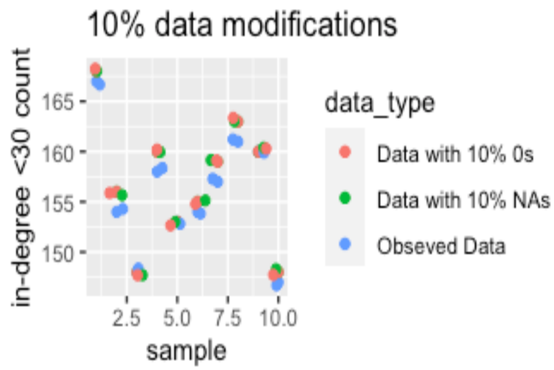


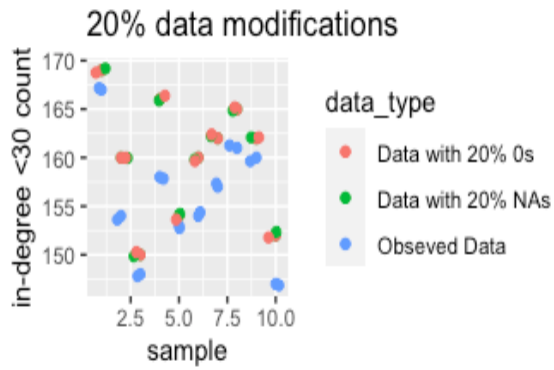


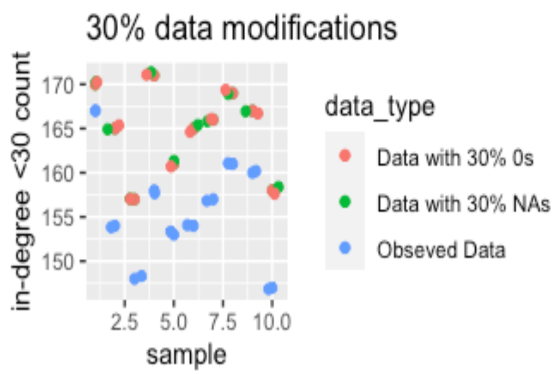


In-degree

The graph below displays the number of actors with an in-degree less than 30. There is a very similar trend as above, indicating that for some actors their popularity appears to be decreasing.

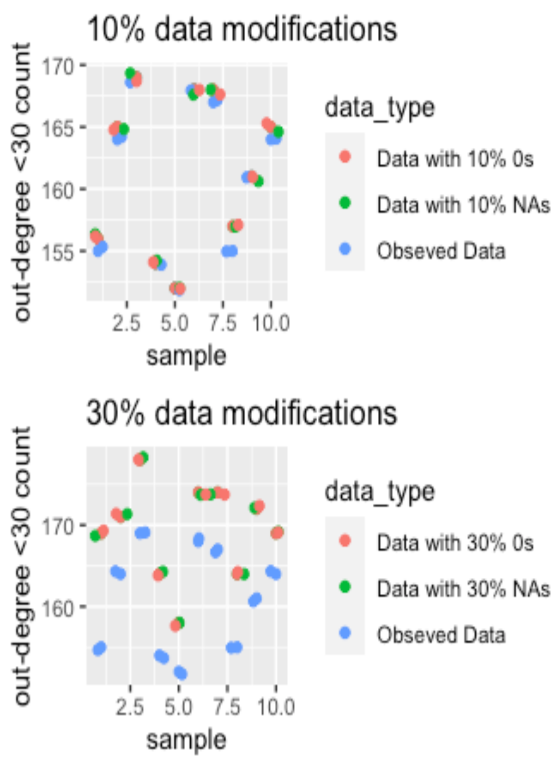






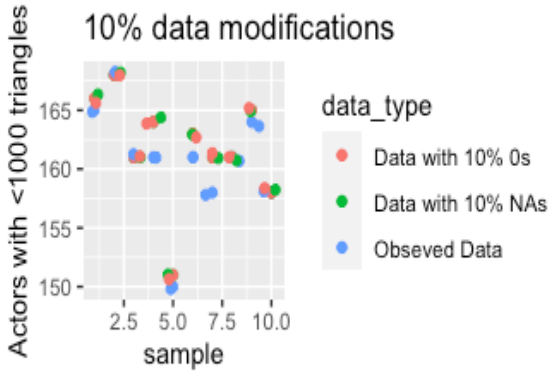
Out-degree

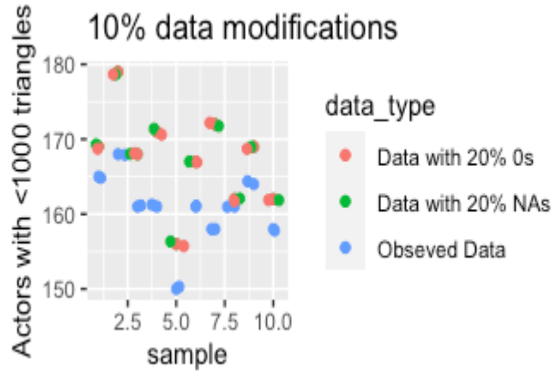
The graph below shows the number of actors with an out-degree less than 30. Again, the trend is very similar to the above two, but this also indicate that for some actors their sociability also appears to be decreasing.

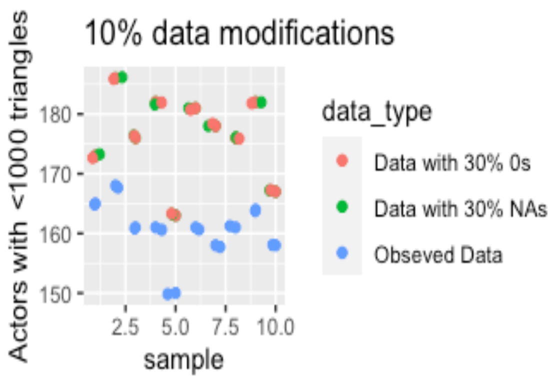


Count Triangles

The graph below represents the number of nodes that have a triangle count of less than 1,000. Thus, as the number of nodes that have a triangle count of less than 1,000 increase this demonstrates that the graph may be becoming more disconnected and have fewer transitive properties. The graph on the right represents the original data, the data replaced 10% with NAs and 10% with 0s. All dots appear to be relatively similar and there are no large differences between the three datasets. However, if we increase the proportion of 0s and NAs in our data to 30% then we see that there is a growing difference between the number of nodes that have a triangles count of less than 1000 compared, where as we increase the number of 0s or increase the amount of corruptness in the data the values of our summary statistics stray further from the values of the full observed data.







Running Social Networks

Using the "amen" package we ran a social networks model on the sampled data, the data that we inserted with NAs, and data that we corrupted by inserting a greater proportion of 0s (Hoff, Fosdick, Volfovsky, & He, 2017). We ran 10,000 iterations on each of the 10 samples, for the full observed data, the data with a proportion of NAs and the data with a proportion of 0s inserted into the data. Posterior values for our intercept, column coefficient, and row coefficient were calculated. For the various summary statistics mentioned previously such as, in, out and total degree, triangle count, eigen vector centrality, and betweenness, the posterior-predictive values were calculated. The following algorithm was used to compute the posterior predictive values:

1) At the current iteration in the Gibbs sampler update parameters.
2) Sample Z matrix using the values from the updated parameters.
3) Use indicator function to obtain Y matrix from Z matrix.
4) Use the Y matrix to calculate various posterior predictive statistics
   - Number of actors with total degrees <60 (in-degree<30, out-degree <30)
   - Number of actors with a triangle count <1000
   - Difference between the median eigen vector centrality for nations who may be corrupt and all other nations
   - Difference between the median betweenness for nations who may be corrupt and all other nations
5) Repeat step 1 and 2 every 25th iteration in the Gibbs sampler, creating a posterior predictive distribution for that descriptive statistic.

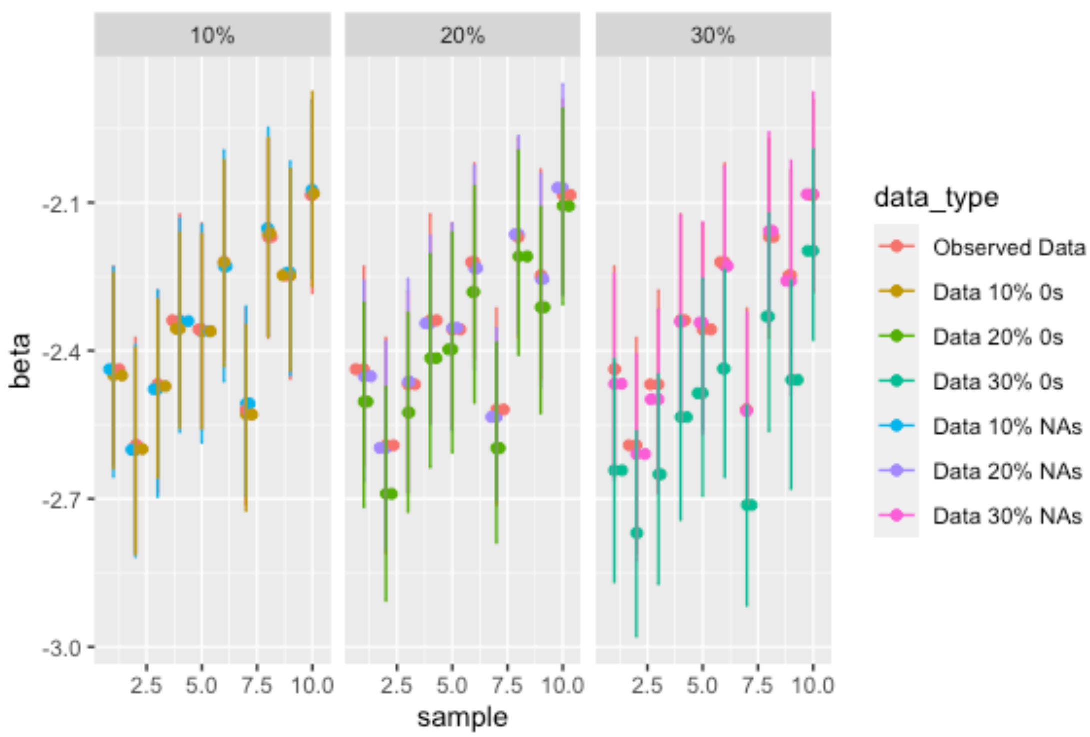
6) Repeat steps 1 through 3 for each observed dataset, each dataset with a proportion of NAs, and each dataset with a proportion of 0s.

Findings

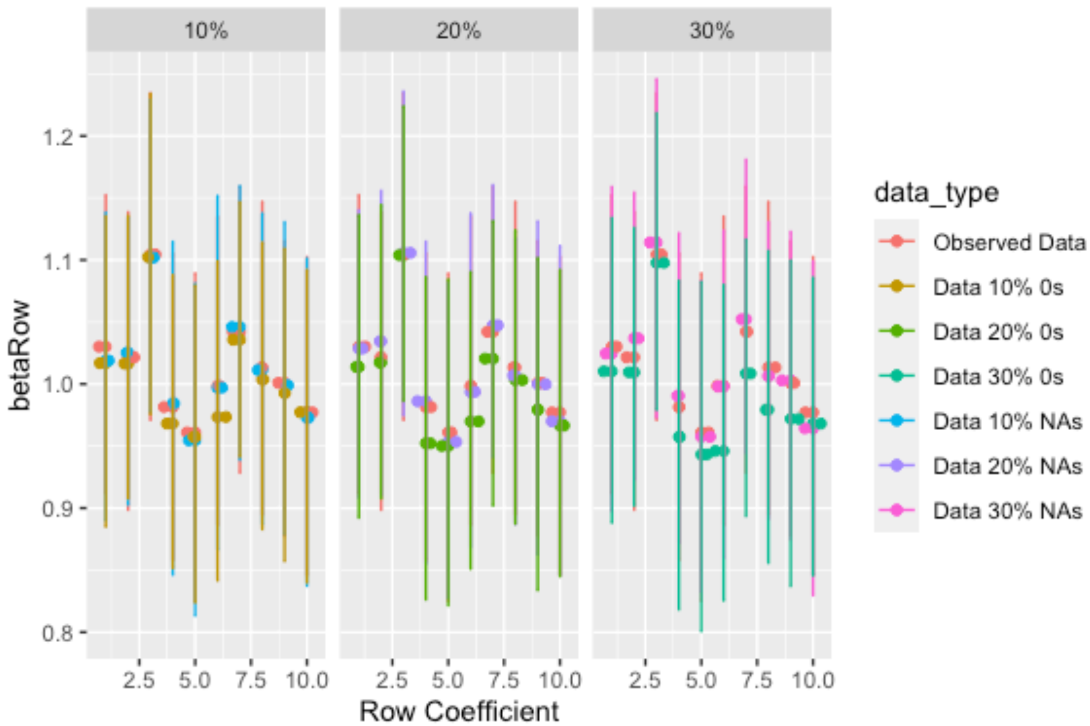
Coefficients

We predict that as the percentage of 0s within the data increase, the intercept will be pulled down and the row and column coefficients will be pulled closer to zero, further from their true value of 1. We also hypothesized that as we increase the number of NAs, it will increasingly become more difficult to accurately estimate the value of the true coefficient but should still be similar. As expected, that data that had a proportion of NAs, though the credible intervals are a wider than the full observed data, the estimates of the coefficients remain very closely aligned with the full data. Providing support for our hypothesis that NAs will not impact the inference but have greater uncertainty around the inference. This trend is continued throughout the following posterior-predictive summary statistics. This implies that as the number of countries who are reporting false arms trade are increasing then the intercept for our model may be pulled down. This trend is present for the column and row means as well. As the number of countries under reporting their data increases then the coefficients for the row and columns get pulled down farther from the estimates of the NA and the observed data.

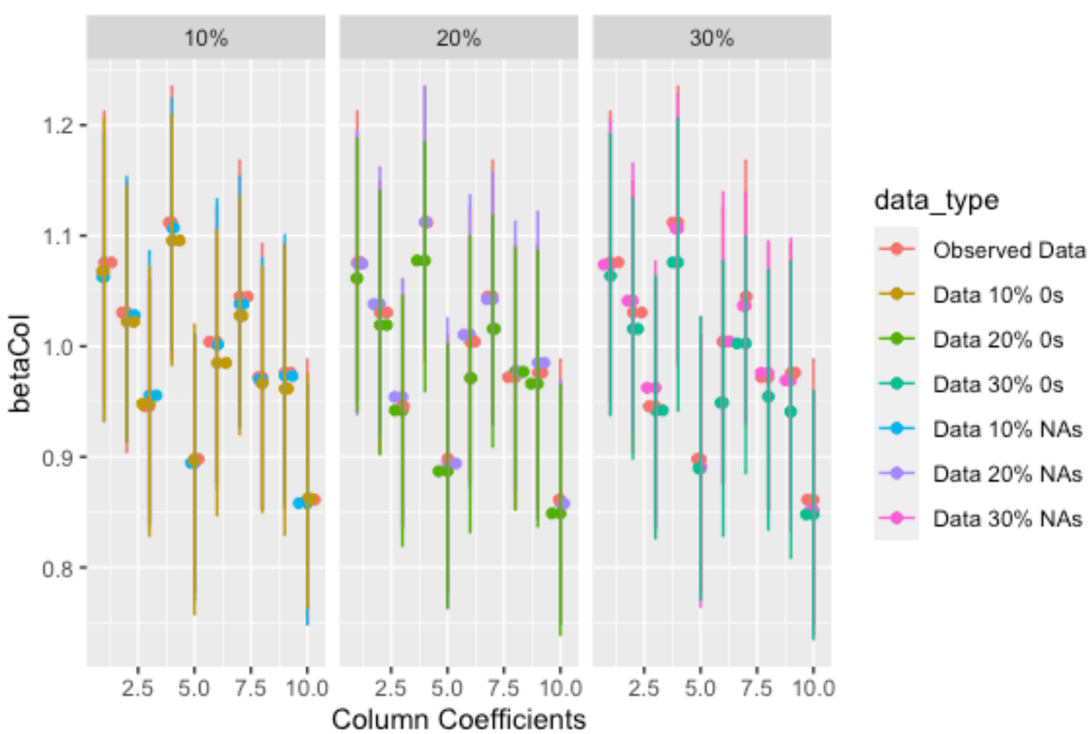
Intercepts



Row Coefficients



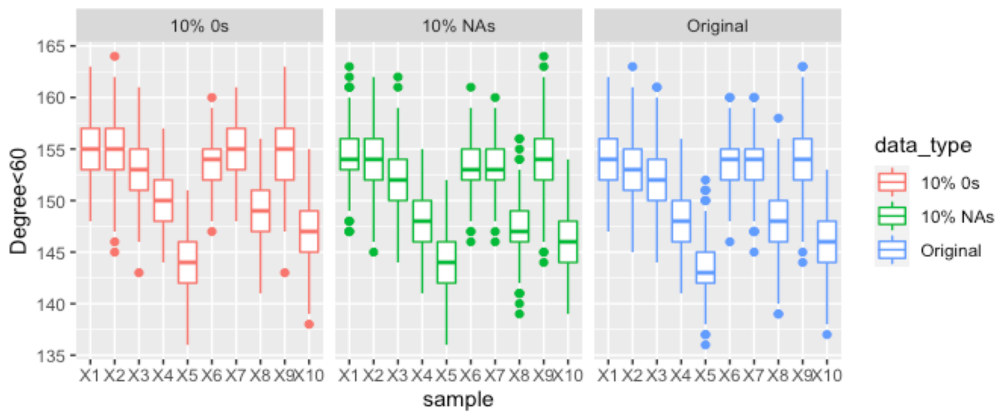
Column Coefficients

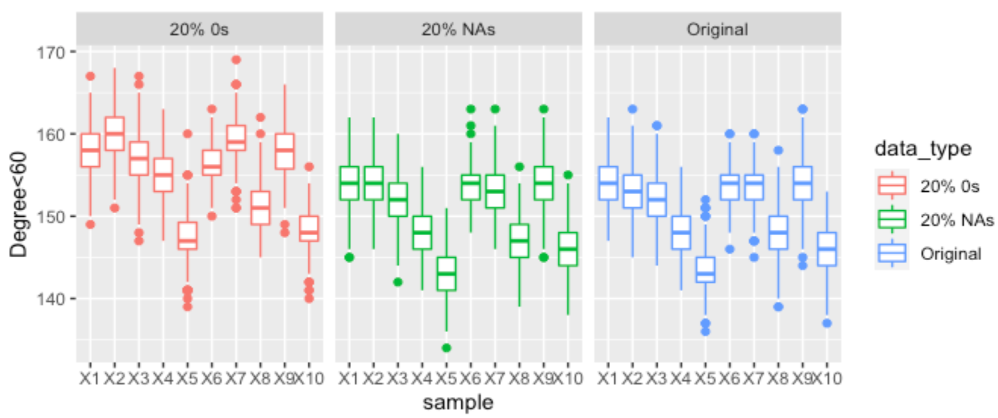


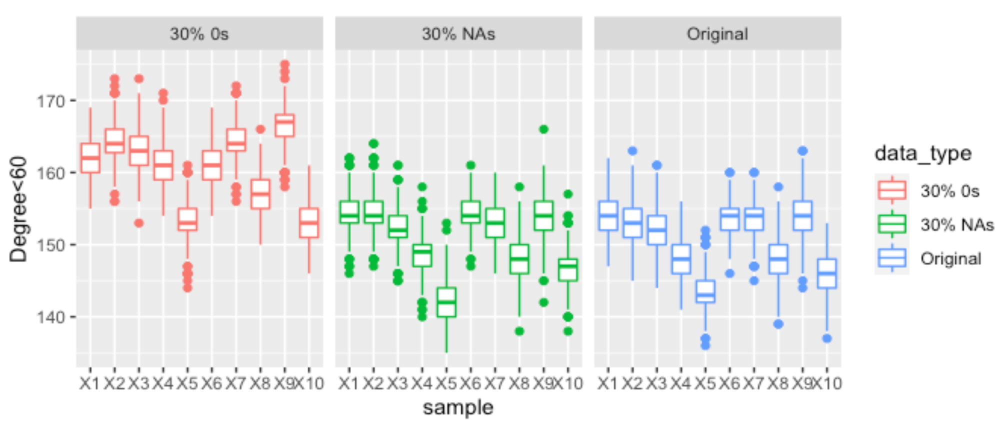
Degree Distribution

All-degree

Here we are looking at the posterior predictive values for the number of actors with a total degree of less than 60. When comparing the three graphs, there is a clear trend as we increase the number of countries who are under reporting their arms trade the number of degrees that are under 60 increases. Looking at the graphs when 30% of the values are corrupted then the boxplots are much higher than those of the observed data and the data with NAs. Make not that even with 30% of the values in the data being replaced with NAs that the degree distribution appears to be very similar to the observed data.

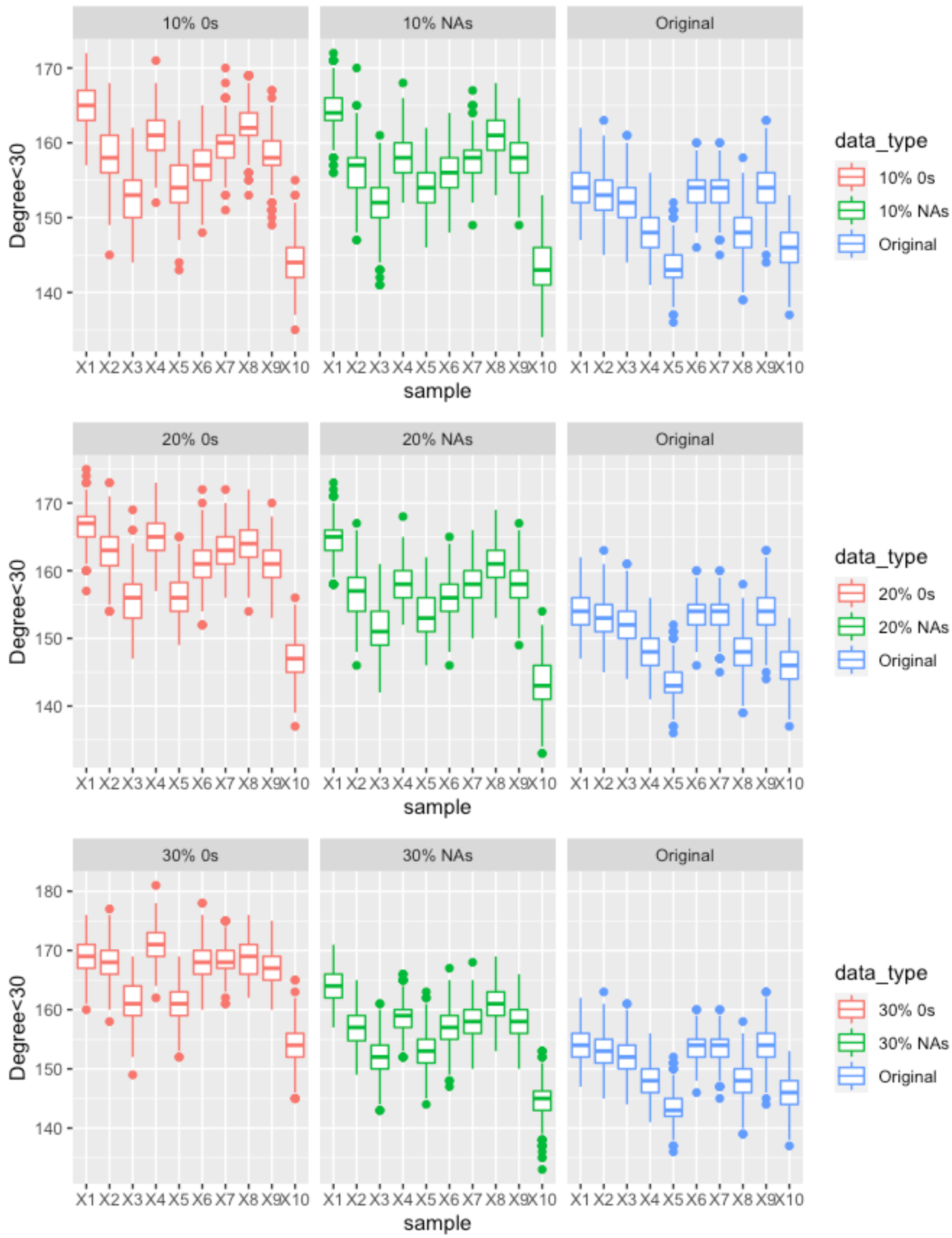






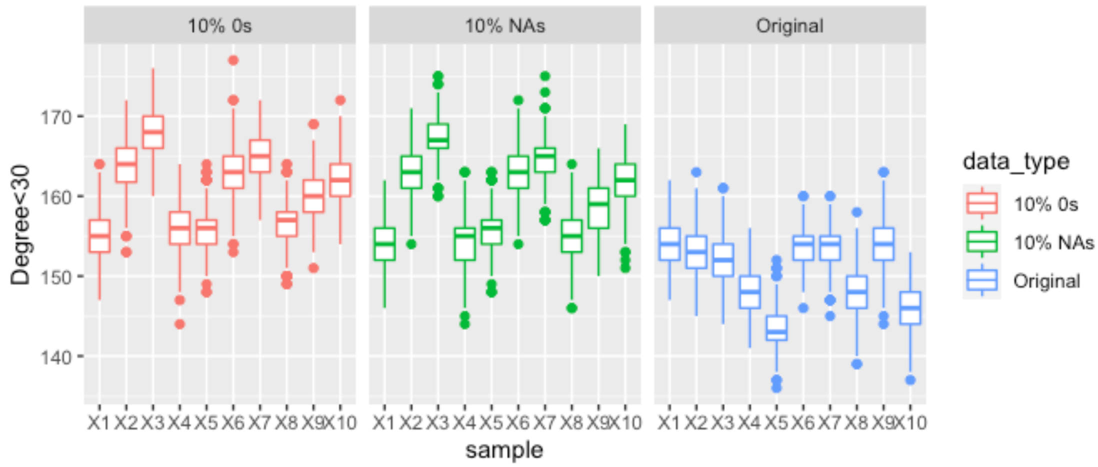
In-degree

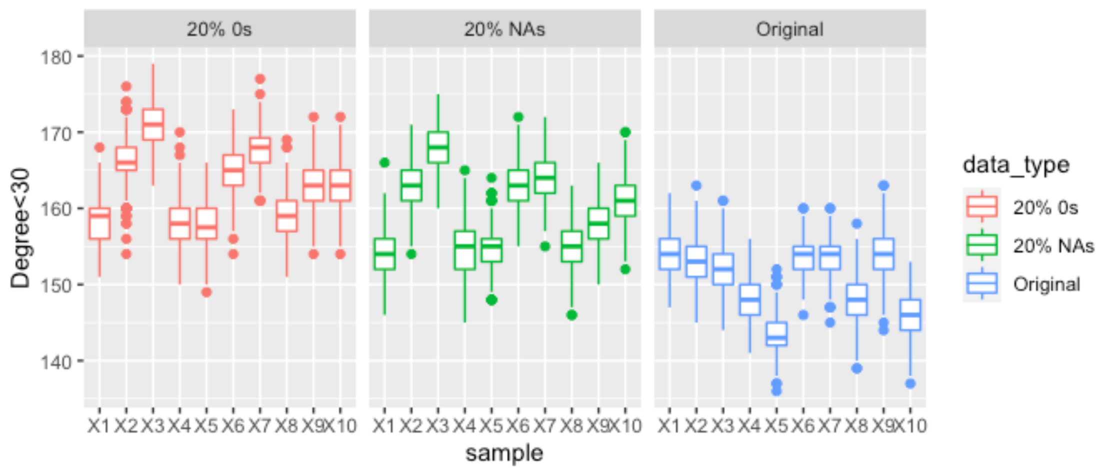
A similar trend as above is shown as the number of degrees under 30 increases, demonstrating that some actors are appearing not as popular the more corrupted is present.

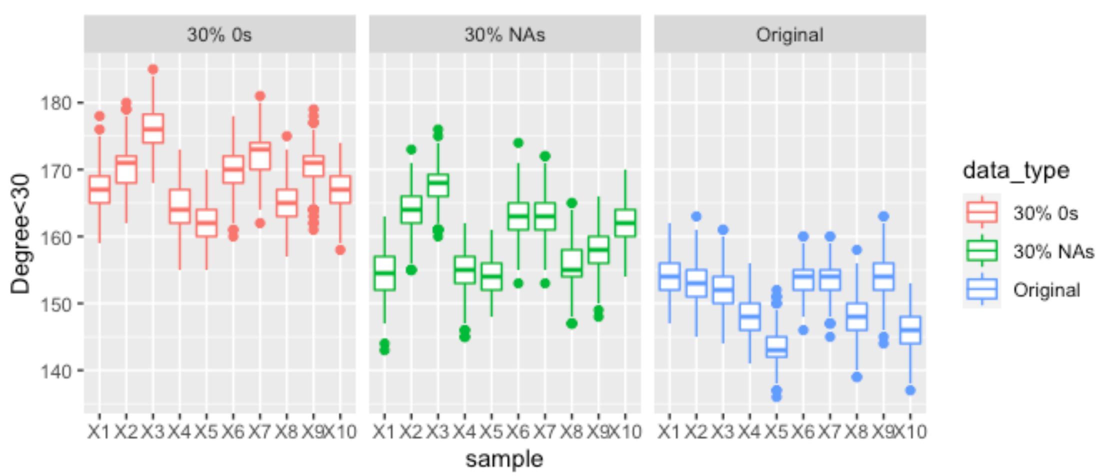


Out-degree

The number of degrees out-degrees under 30 increases showing that some actors are appearing less sociable as the amount of corrupted data increases.

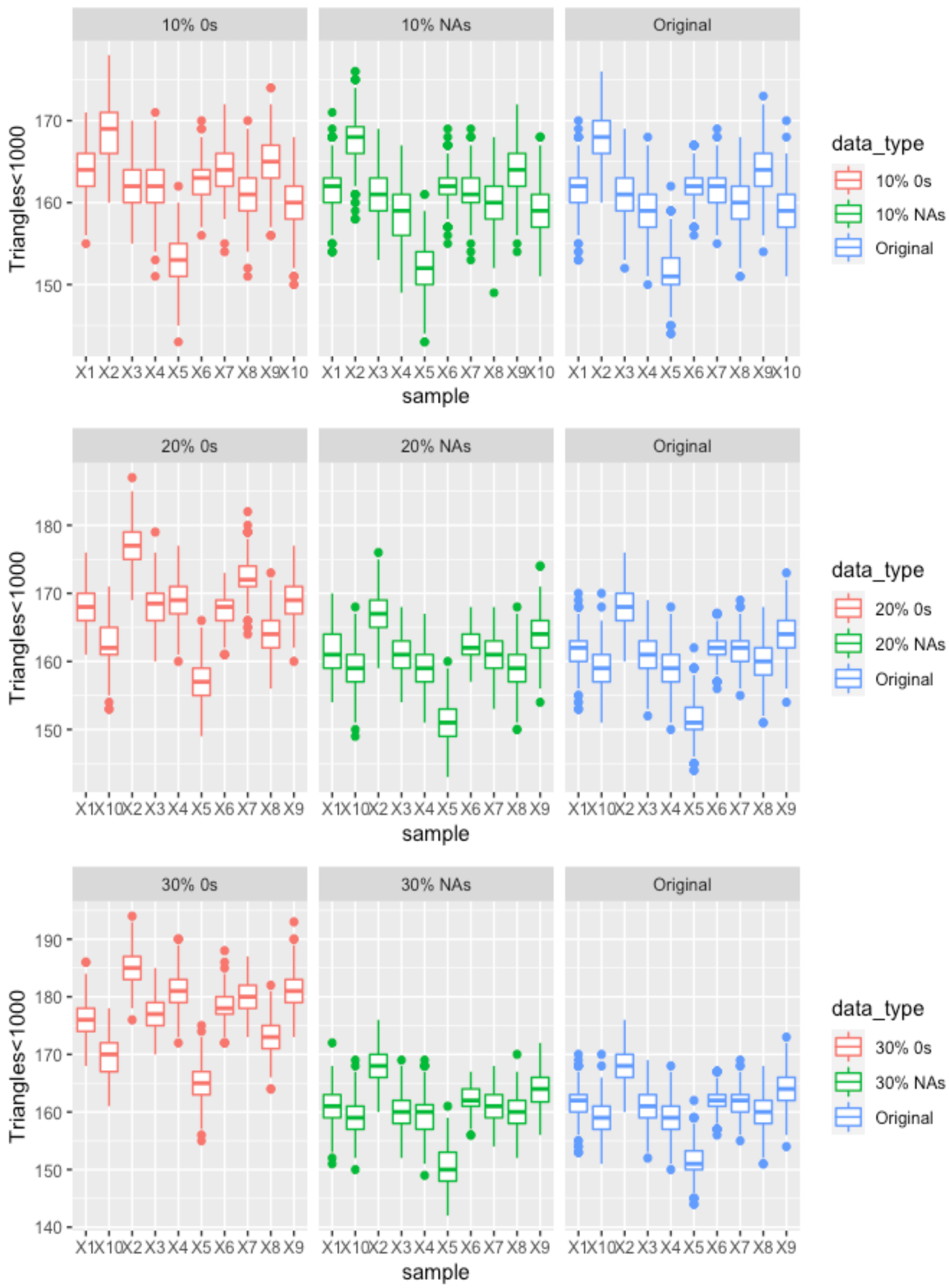






Counting Triangles

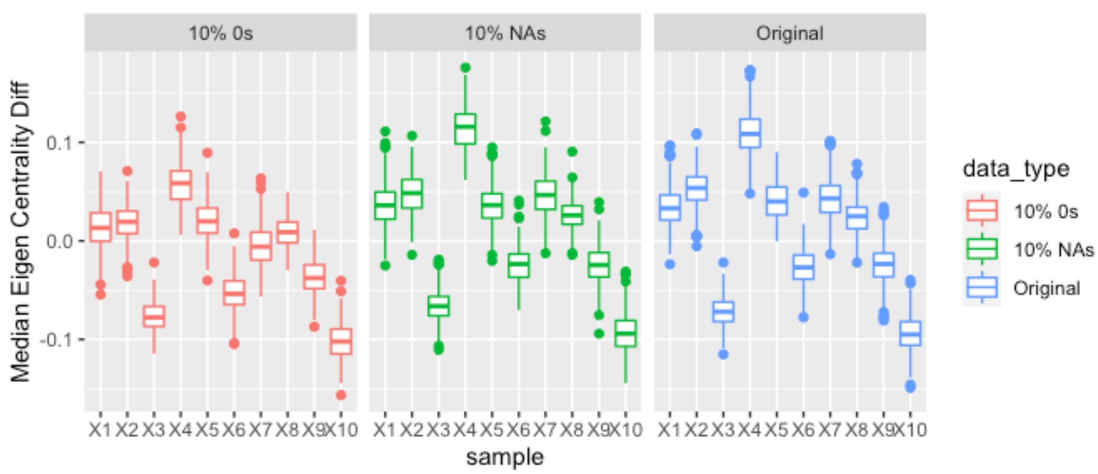
Analyzing the graphs over the varying proportions of NAs and 0s, we see that as the number of 0s increases there are a larger portion of nodes that are a member of less than 1000 triangles. This demonstrated that the graph grows increasingly less connected and might have altered some of the clusters originally present. It's also important to note that the data that had a proportion replaced with NAs is very similar to the original distribution, this is in line with our initial hypotheses. Since the missing values are being imputed from the generative model, we would expect their inference to be close to the original samples.

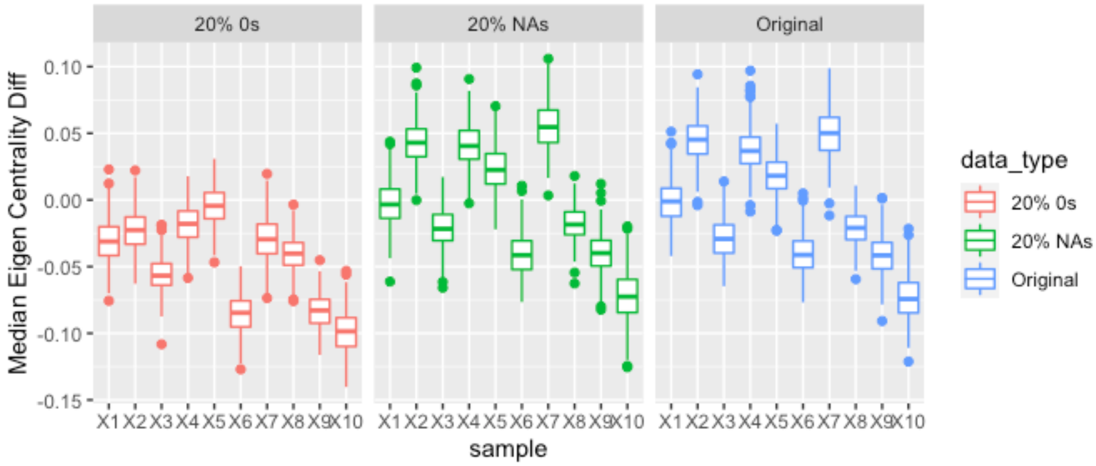


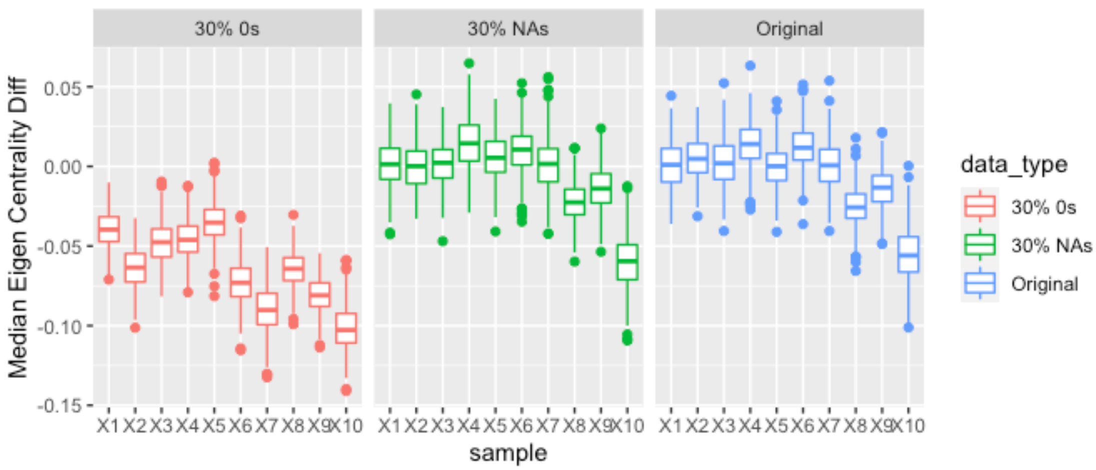
Eigen Vector Centrality

For nodes that are underreporting their arms, they will have fewer relations and thus appear not as important or influential to the network as they might be. Thus, nodes that are underreporting may tend to have a lower eigen vector centrality than nodes that are reporting all their arms trades. To get at this difference we took the three types of data we had: the observed data, the data with a proportion of NAs, and then the data with a proportion of 0s. Then we looked at the posterior predictive eigen vector centrality for the nodes at each iteration in the Gibbs sampler. The median eigen vector centrality for the nodes that were considered to be underreporting was calculated and then the median eigen vector centrality for the rest of the nodes was calculated. The difference of these two values was then computed. Thus, for the observed data and for the data with a proportion of NAs we would not expect there to be any difference for the distributions of the difference in medians to be centered at 0. However, for the data where we inserted a proportion of 0s, we would expect the median eigen vector centrality for the nodes that were underreporting their arms trade to be less than the eigen vector centrality for the rest of the nodes who are accurately reporting their data.

This result was observed, and as we increased the proportion of nodes that are underreporting, we see that the difference between the median eigen vector centrality of the nodes that are underreporting also decreases and the nodes that were reported accurately increased.

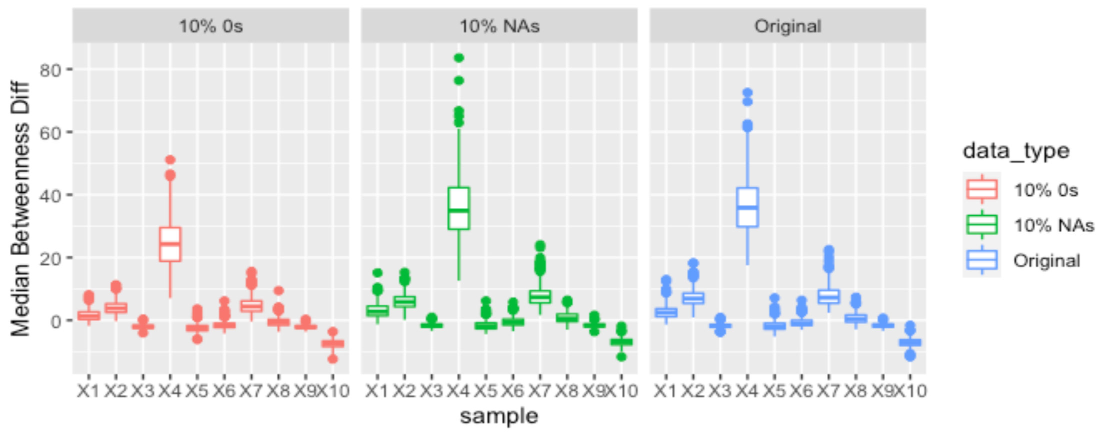


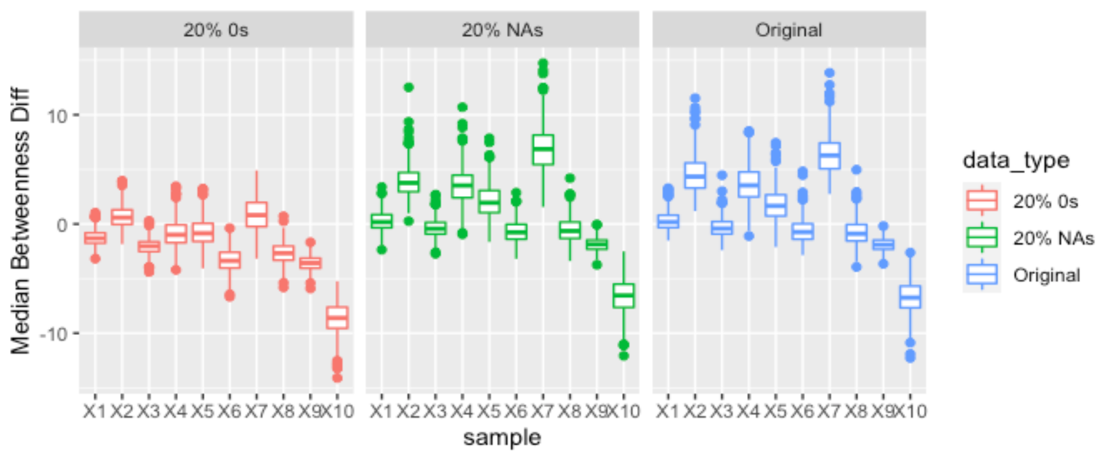

20% 0s
20% NAs
Original
Median Eigen Centrality Diff
0.10
0.05
0.00
-0.05
-0.10
-0.15
X1 X2 X3 X4 X5 X6 X7 X8 X9 X10
sample
data_type
20% 0s
20% NAs
Original

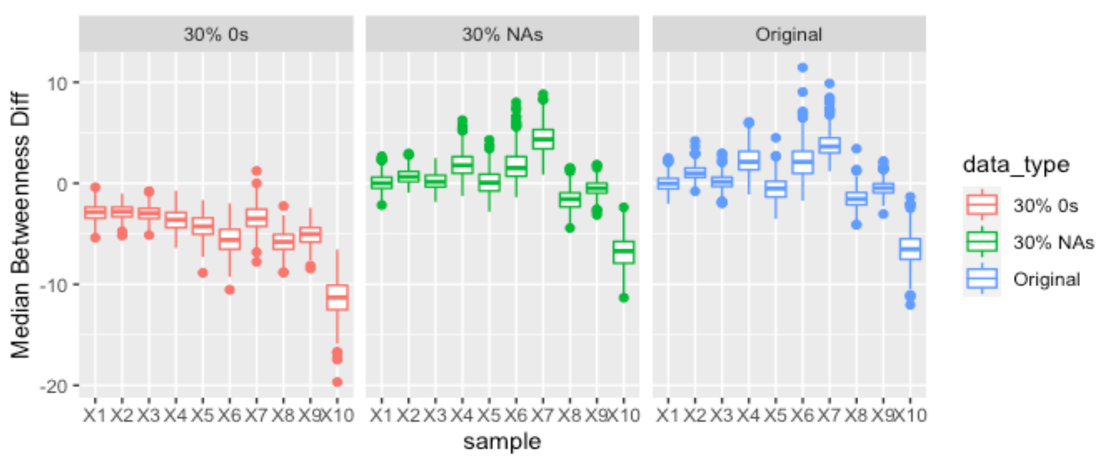

30% 0s
30% NAs
Original
Median Eigen Centrality Diff
0.05
0.00
-0.05
-0.10
-0.15
X1 X2 X3 X4 X5 X6 X7 X8 X9 X10
sample
data_type
30% 0s
30% NAs
Original

Betweenness

For betweenness we hypothesized that the nodes that were not reporting their full arms trade would have a lower betweenness score than countries that were reporting their full arms trade. Similar to eigen vector centrality we would expect for nodes that underreported their arms, their median betweenness score would be less than the nodes who reported their full arms trading. This trend was observed, and as the proportion of 0s was increased the difference between the median betweenness of the nodes underreporting and the nodes that were accurately reporting increased.







Hypothesis Testing

Though the posterior predictive distributions above show a visual difference in the distributions between the three data types, we wanted to quantify this difference. When carrying out analyses with small arms data that may be corrupted, the main question that we are trying to address is are the data that have been replaced with a percentage of 0s coming from the same distribution as the original sampled data, in other words, is Y_0 coming from the same distribution as Y_{obs}? We designed a hypothesis test to examine this question, the following method was taken:

1) To get the null distribution, we sampled 1,000 data sets from the original generative model.
2) The test statistics for each distribution was then calculated (number of actors with total-degrees less than 60, number of actors with in-degree less than 30, number of actors with out-degree less than 30, number of actors with a triangle count of less than 1,000).
3) We then examined the posterior predictive distribution of Y^{obs}, Y^{NA}, and Y^{0} for each of the 4 summary statistics. The following questions are:

a. In this resampled graph, do I have more nodes with a degree fewer than 60, than one would expect under the null distribution?

b. In this resampled graph, do I have more nodes with a triangle count fewer than 1,000, than one would expect under the null distribution at each iteration?

4) Then the posterior predictive p-value is calculated with respect to the empirical distribution of the count of triangles that equal 0 or 1. $P(T > T^{obs}|thenull)$.

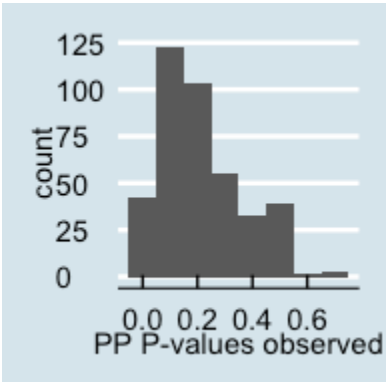
The posterior predictive p-value represents a measure of discrepancy between the null value and the sampled data.

Hypothesis Testing Findings

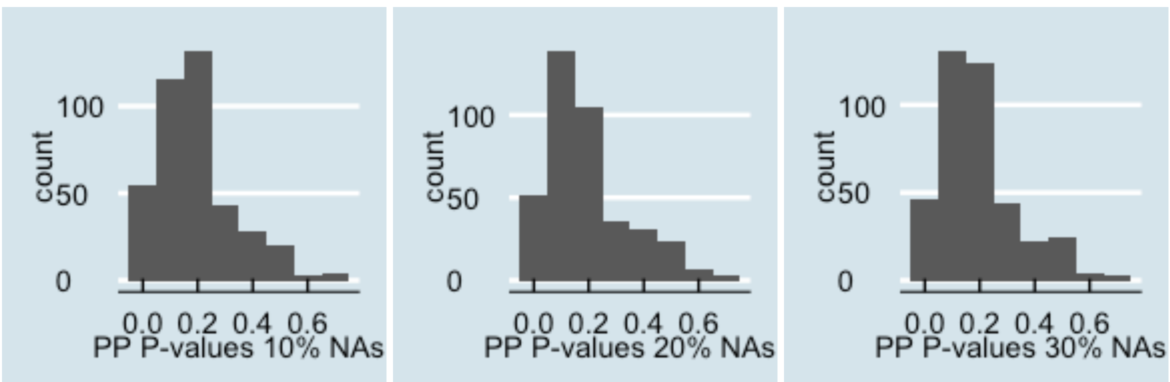
Degrees

After completing the above algorithm for computing posterior predictive p-values these values are plotted in order to see the distribution of the posterior predictive p-values. If we initially consider the posterior predictive p-value distribution for the original data, then we see that only a small number of the samples are getting rejected (are less than .05). A similar trend is present when looking at the data that originally had a proportion of values replaced with NAs. This is what one should expect as these missing values are being imputed from the generative model. This demonstrates that even if there are high levels of corruption in the data that replacing theses values with NAs will allow the inference to remain relatively unaffected when considering the number of total-degrees under 60.

Posterior Predictive P-values for the observed data

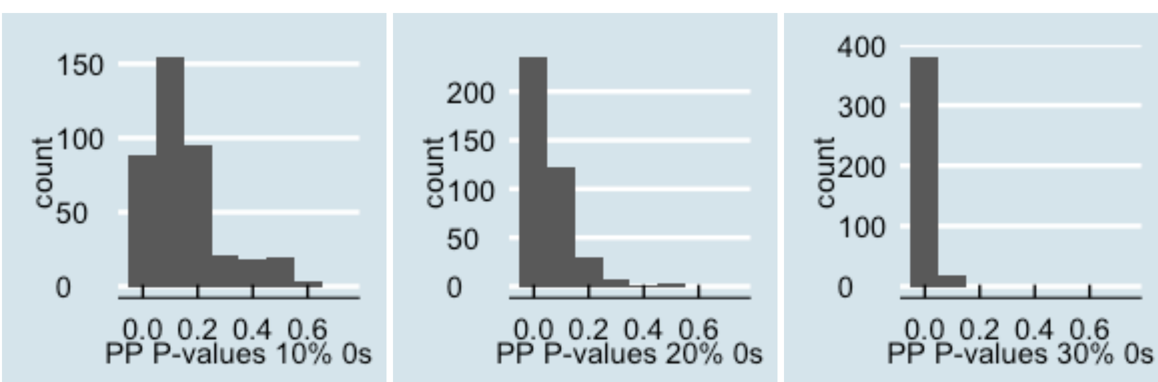


Posterior Predictive P-values for data with a proportion of NAs



Upon examining the data that has a proportion of zeros, we see that as the amount of corrupted data increases the posterior predictive p-values all become less than .05 signifying that for all values in the posterior predictive distribution, the distribution of the number of actors with degrees under 60 is different from the null distribution. Similar trends are found for in-degree and out-degree descriptive statistics.

Posterior Predictive P-values for data with a proportion of 0s

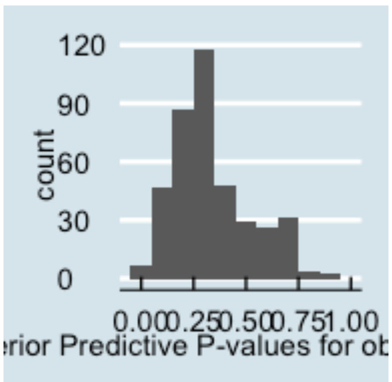


The graphs above present the results of one sampled dataset out of the 10 that we initially sampled from the generative model, but the trend is the same throughout all 10 samples.

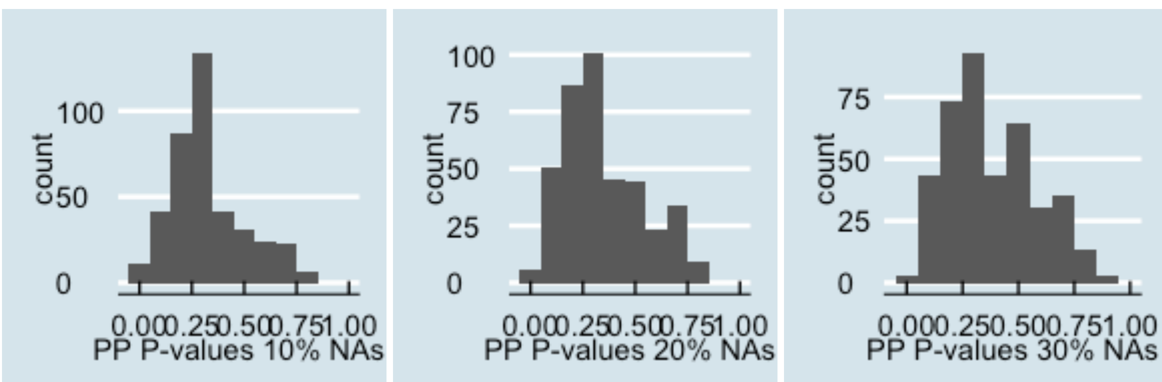
Count Triangles

A similar trend as above is found for counting triangles. The posterior predictive p-values generated from the data with a proportion of the values replaced with NAs is very similar to the distribution of p-values form the full observed data. However, increasing the proportion of corrupted data lead to the posterior predictive distribution of actors with a triangle count of less than 1000 that is different from the null distribution, demonstrated by almost all of the p-values being less than .05.

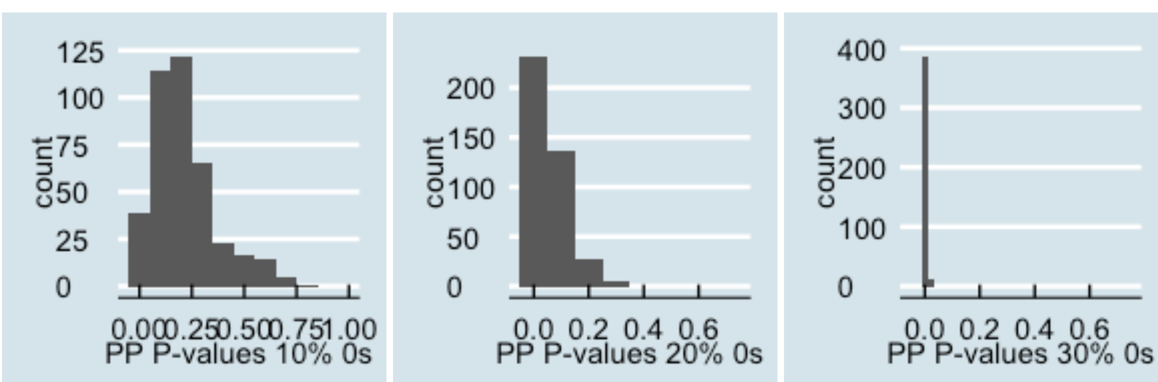
Posterior Predictive P-values for the observed data



Posterior Predictive P-values for data with a proportion of NAs



Posterior Predictive P-values for data with a proportion of 0s



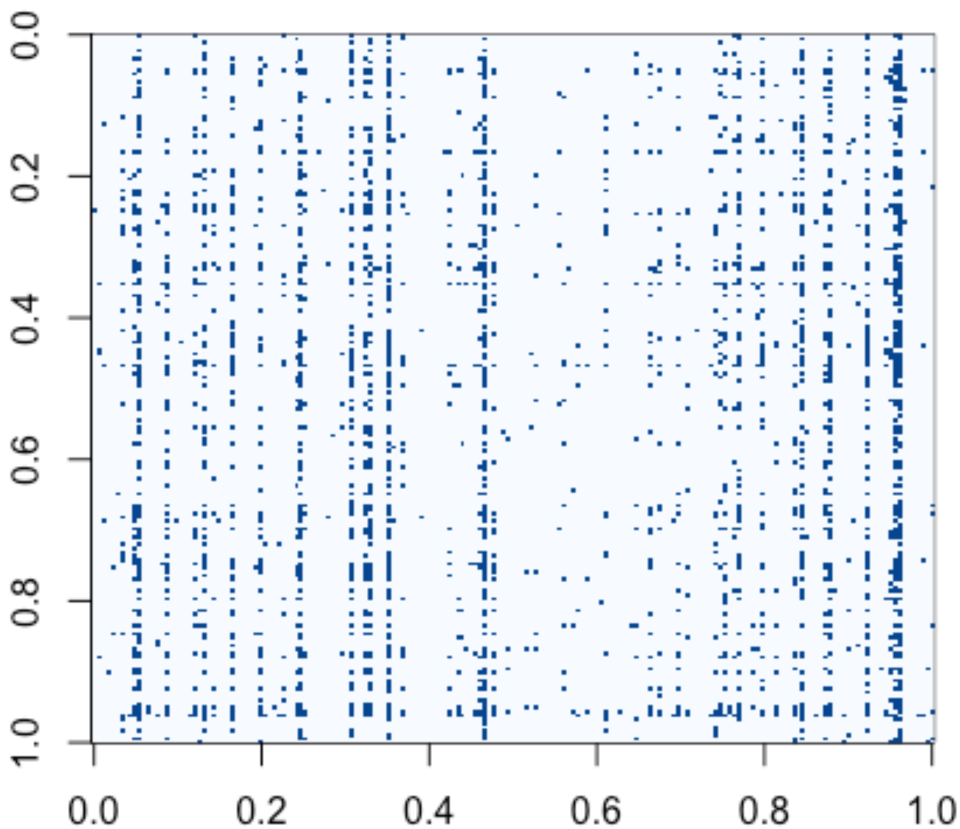
As we can clearly see from the posterior predictive p-values, the corrupted data drastically impacts the posterior predictive distribution for the above descriptive statistics. When conducting analysis with possibly corrupted data, and analyzing theses aspects of the network it is important to consider how the distribution may look if the corrupted observations were replaced with NAs.

Chapter 4

Application to Arms Data

Norwegian Initiative on Small Arms Trade

The Norwegian Initiative on Small Arms Trade is a database containing dyadic information on small arms trades from 1962 to 2015. The data includes the seller, the buyer, the type of weapon transferred, the year of the transfer, the total value of the arms purchase, and the source of the information. This data is both directed and weighted, where the direction is from the seller to the buyer and the weight is the number of weapons sold. However, mimicking the analysis above, we transformed this data into a binary network where, $Y_{i,j} = 1$, represents a transfer of arms from country i to country j. Below is an adjacency matrix of arms trade from the year 1997:



We can notice from this socio-matrix a few important aspects. First that there a few countries who seem very popular but not as many nations who sand out as being very sociable. Also there are around 10% of entries in the matrix are 1s.

Below is also an expert form the data from the year of 1997, showing the initial format of the data and the information provided. Note that prio_weapons_code refers to the type of weapons that we transferred. Weapons within the 100 range are light weapons, and increasing numbers represent an increase in destructive caliber of the weapon where in the 400s this includes arms such as landmines and explosives, while most of the transfers in the 500s represent ammunition and machine parts.

Below is also an expert form the data from the year of 1997, showing the initial format of the data and the information provided.

seller	year	value	prio_weapons_code	buyer	data_source
Afghanistan	1997	1979	210	Czech Republic	6
Andorra	1997	2780	223	France	6
Andorra	1997	8639	223	Spain	6
Andorra	1997	13668	227	Spain	6
Argentina	1997	995	227	Austria	6
Argentina	1997	39516	210	Bolivia	6
Argentina	1997	319133	210	Brazil	6
Argentina	1997	22533	417	Brazil	6
Argentina	1997	8054	210	Canada	6
Argentina	1997	30259	210	Chile	6

This data may be corrupted and thus as we saw in the previous section, doing analysis on this data without dealing the corruption could greatly impact any statistical inference we perform. If there is an underlying generative model then by replacing the values of the countries that are under reporting their arms trade with NAs, we would be able to impute these values from the true generative model. If the countries chosen are indeed engaging in illicit arms trading, then we would expect to see a difference in our inference if we ran the data as it is currently versus when we replace the possibly corrupted data with unknowns.

We identified just under 30 nations that we thought may have been engaging in illicit arms trading during the time period of the 1990s to 2005. These nations were chosen based on previous account of illicit arms trading (Chile, Niger, Afghanistan), were engaged in conflict or had been embargoed (Bosnia and Herzegovina, Israel, Pakistan, India), countries that are boarding an embargoed nation (Somalia, Ethiopia), and countries we believe may not reporting their arms trade due to political or profit based motivations (Russia, China, India, Venezuela). Below is a map where the countries we think are trading legally are in blue and the countries that we think may be engaging in illicit arms trading are in red.

After identifying these countries, we then made copies of the adjacency matrixes and for a trade that occurred between two nations that we identified as possibly under reporting, an NA was placed in that entry of the matrix.

In order to test our hypothesis, we ran the same analysis as above but the original arms data. The model was set up in the following way:

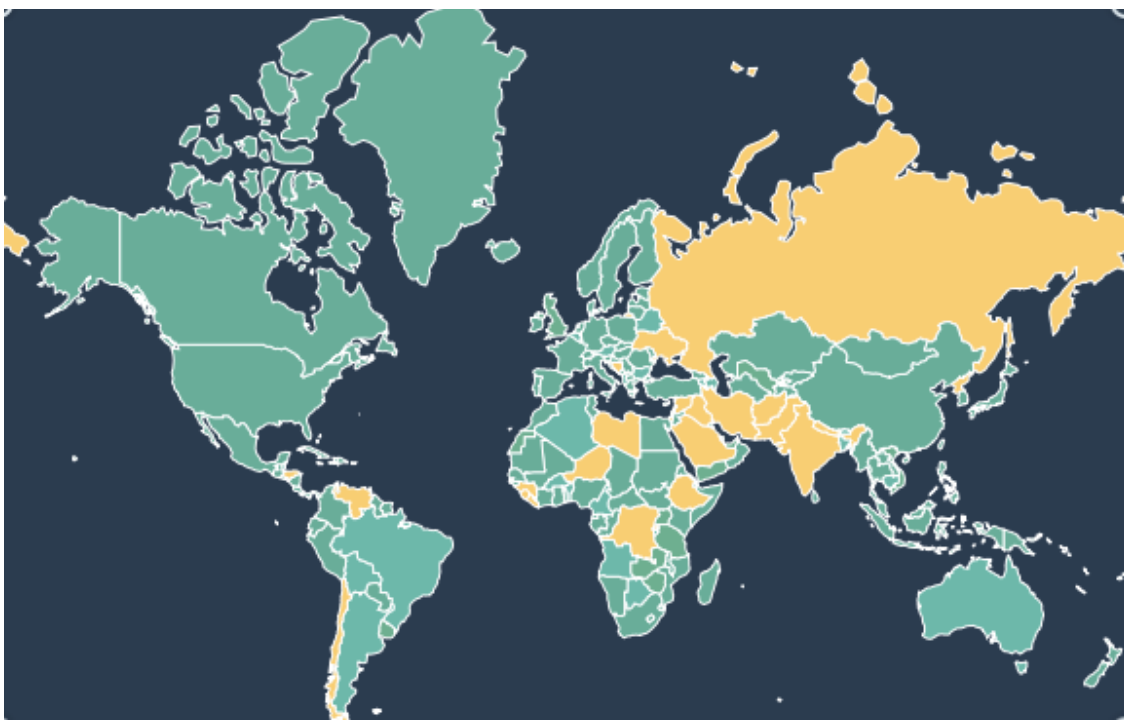
$z_{i,j} = \beta^T x_{i,j} + a_i + b_j + \gamma_{i,j}$

$y_{i,j} = 1(z_{i,j} > 0)$

Where Y is s binary matrix, where $y_{j,j} = 1$ represents that country i sold arms to country j. This analysis was also divided up by year where each Y socio-matrix represented a all the reported small arms trade from year t, for $t \in (1995,2005)$.

Two covariates were included in the model. The nodal covariate was the previous year's standardized log(GDP) (at time t-1). This was used as both a row and column covariate. The other covariate was a dyadic covariate and was the previous year's arms trade, arms trade at (t-1).

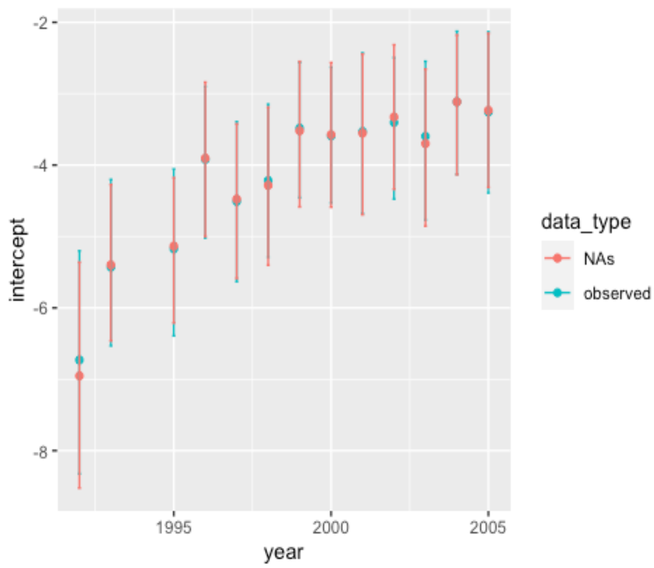
Below is a map where the countries we think are trading legally are in blue and the countries that we think may be engaging in illicit arms trading are in orange.



Findings

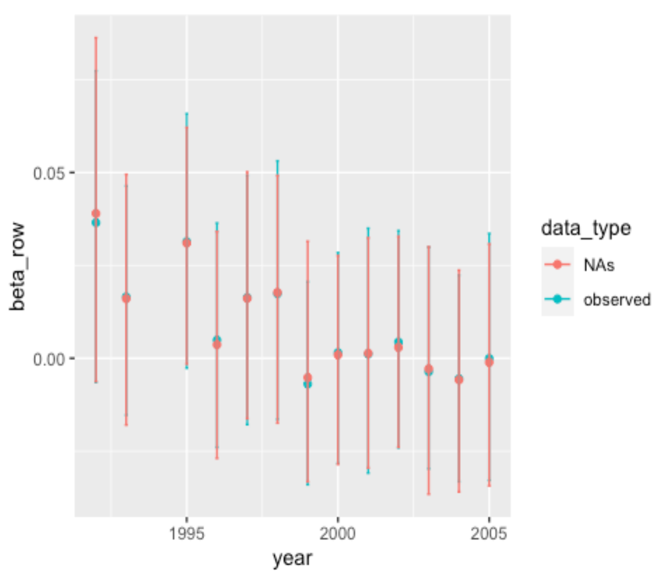
Beta (intercept)

When observing the intercept, the row and column coefficients for the nodal variable, and the observed data and the data with NAs inserted in appears very similar across all years.



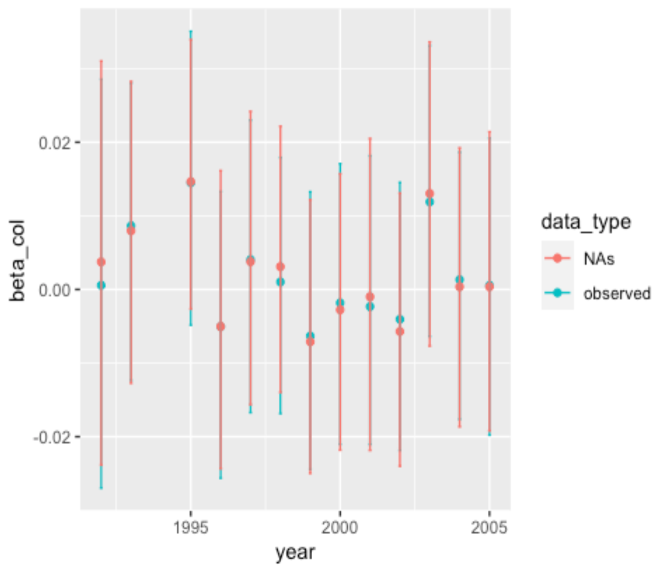
Data type	Mean Beta	sd Beta	Median beta
NAs	-4.166469	1.097381	-3.698587
observed	-4.150008	1.057090	-3.596627

Beta (row)



Data type	Mean Beta	sd Beta	Median beta
NAs	0.0087514	0.0141731	0.0029268
observed	0.0087686	0.0138486	0.0043808

Beta (column)



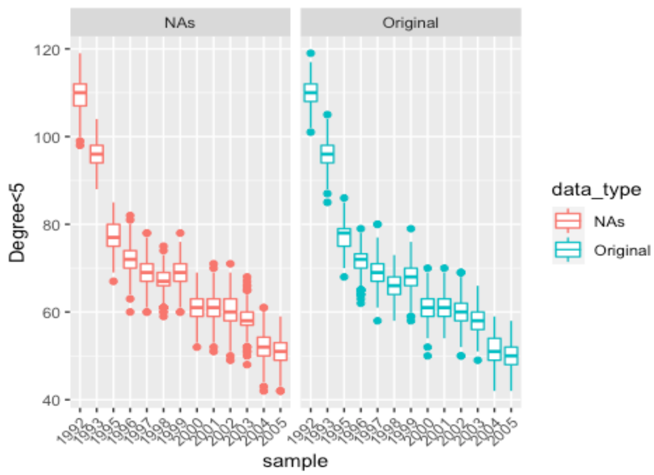
Data type	Mean Beta	sd Beta	Median beta
NAs	0.0019543	0.0067890	0.0003718
observed	0.0017676	0.0064314	0.0005590

Dyadic Coefficient

data_type	Mean_Beta	sd_Beta	Median_beta
NAs	1.919837	0.0964773	1.914063
observed	1.914881	0.0895733	1.906568

Total Degrees

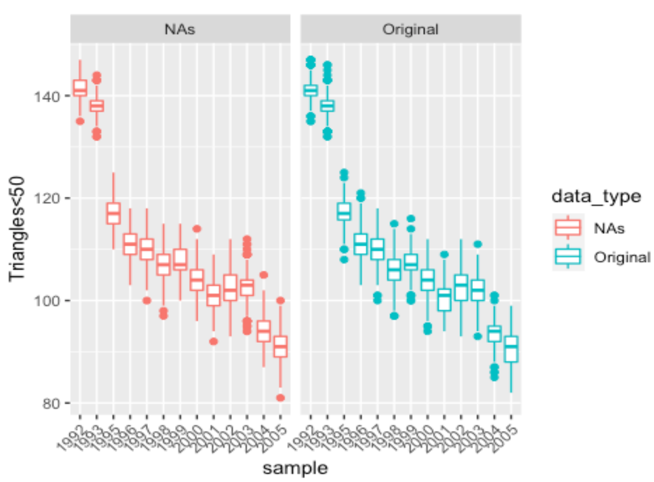
For the total, in and out degrees the distribution from the original data and the data replaced with NAs look very similar. While the original data has an ever so smaller number of the number of actors with



Data type	degree all mean	degree all median	degree all sd
NAs	69.49500	66	16.49296
Original	69.17038	65	16.68179

Triangle Count

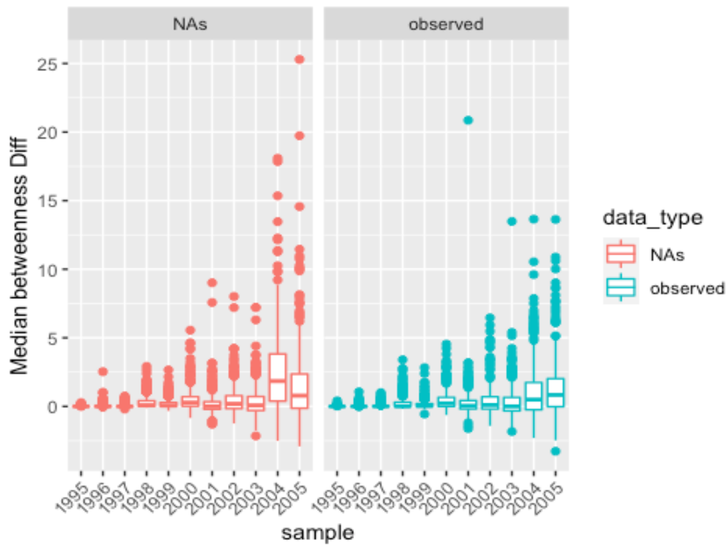
The number of actors with triangles that are less than 50 is again very similar across both groups.



Data type	Triangle mean	Triangles median	Triangles sd
NAs	109.6958	106	14.62488
Oiginal	109.5444	106	14.71470

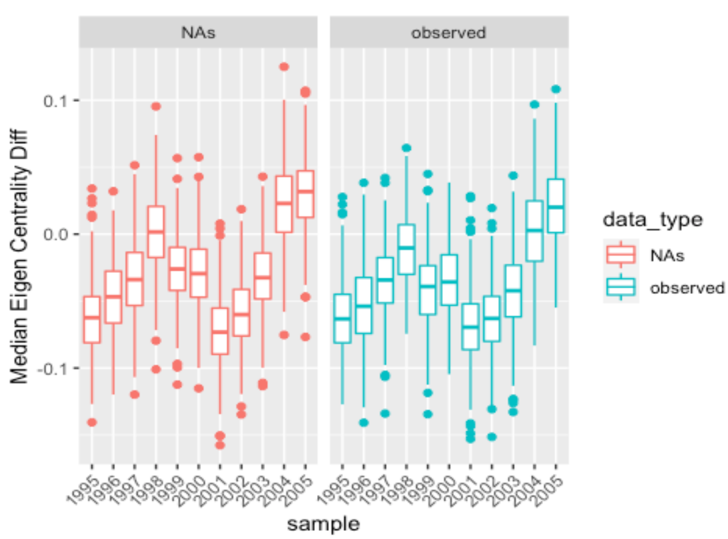
Betweenness

There appears to be a greater variance around the estimates for the data where NAs were inserted. This is what we would expect as though this data is coming from the generative model there is less certainty around the estimates. We also can observe that there appear to be a greater number of outliers for the data with NAs. While the uncertainty has increased and some more extreme outliers for some of the years, the actual estimates appear to be very similar.



Eigen Vector Centrality

We can see from difference in the medians of eigen vector centrality of the countries we believe may be illicitly trading small arms and countries we think are reporting the accurate levels, we see that for some years there is some difference in the eigen vector centrality between the groups. This may indicate some alteration in the centrality due to illicit trading.



Chapter 5

Discussion and Conclusion

The presence of corrupted or underreported data can drastically impact our inference, especially as the proportion of corrupted data increases. Networks can appear less connected, nodes can appear to be less influential, and predictors can appear to be less important. The goals of this project were to set up a method for researchers to notice when the arms trade data they have may be corrupt. Our methods section indicates that signs of corruption may include things such as the nations suspected of illicitly trading may have a less betweenness and less centrality in the network than nations who are not illicitly trading. Data that is corrupted may also report a greater proportion of actors with degree and triangles under a certain threshold than the true data or the data with the corrupt values replaced with NAs. This is due to the fact that when nations fail to report the true number of arms they are trading, this decreases their number of degrees (in, out and total) and their connectedness to the rest of the network. This is particularly concerning when we are studying what nations are most influential and important to the arms trade and international relations as a whole, as our initial conclusions may not be based on correct data. This analysis also indicates that an augmentation of the proportion of countries who are illicitly trading leads to greater discrepancies in our inference. However, it may be more difficult to definitively say if nations are illicitly trading arms if that proportion lies around 10-15 percent.

The application to real data demonstrates that these countries may not be engaging in illicit arms trade or not all of them are engaging in illicit arms trade. While many of the posterior predictive statistics such as the number of actors with triangles less than 50 and the number of actors with total and in degrees less than 5 are similar across the reported data and the data we replaced with NAs for some years; However, many years these values have some discrepancies and thus many indicate that some illicit trading occurred. This difference is too small to draw any definitive conclusion that the thirty countries choses may be engaging in illicit arms trading but provides us with a note of caution when using this data. We also see that for the eigen vector centrality and the betweenness scores, the countries we believe to be trading illicitly appear to be less central and have less betweenness than when we observe the data replaced with NAs. Thus, when using this data for future analyses it is important to proceed with caution or consider using the data with countries we think are illicitly trading replaced with NAs.

There is a lot of future work to be done within this area of research. In particular dealing with circumstances of when nations are illegally trading to other nations but *legally* trading to other nations that my also illegally trading. Thus, this analysis assumes that if a country is trading illicitly then it is trading illicitly with all other countries that are trading illicitly. Further work would need to be done to see the extent of data of this context.

A more daunting project to take on when dealing with data of this sort if the possibility that this data would not be missing at random. There is a systematic relationship between whether a nation is not reporting how much arms they trade and the number of arms they trade. This analysis assumes that data would be missing at random, which allows these values to be imputed form the generative model.

Arms trade data is used to make important policy and military decisions, and the use of incorrect data can have detrimental impacts for populations and nations. Being able to identify when information may be misleading is essential for international security. Using the methods outlined above one can see where these discrepancies may occur, and if policies such as arms embargos are being effective or being undermined. Failing to account for illicitly traded arms is hindering our ability to understand the effectiveness of arms embargos, threats to international security, and international relations dynamics as a whole. The hope is that this work opens up other opportunities and ideas to predict and determine with whom and when is illicit arms trading occurring.

References

Arsovska, J. (2014). Illicit Firearms Market in Europe and Beyond. European Journal on Criminal Policy and Research, 295–305.

Bichler, G., & Malm, A. (2013). Small arms, big guns: a dynamic model of illicit market opportunity. Global Crime, 261-286.

Brzoska, M., & Lopez, G. A. (2009). Putting Teeth in the Tiger: Improving the Effectiveness of Arms Embargoes. Bingley, United Kingdom: Emerald Group Publishing.

Buuren, S. v. (1999). Flexible multivariate imputation by MICE. Leiden: Flexible multivariate imputation by MICE. DellaVigna, Stefano, & Ferrara, E. L. (2010). Detecting Illegal Arms Trade. American Economic Journal: Economic Policy, 26-57.

Hoff, P. (2002). Latent Space Approaches to Network Analysis. Journal of the American Statistical Association, 1090-1098.

Hoff, P., & Ward, M. (2004). Modeling Dependencies in International Relations Networks. Political Analysis, 160–175.

Hoff, P., Fosdick, B., Volfovsky, A., & He, Y. (2017, 5 25). Additive and Multiplicative Effects Models for Networks and Relational Data.

Phythian, M. (2000). The illicit arms trade: Cold War and Post-Cold War. Crime, Law & Social Change, 1-52. Rothe, D. L., & Collins, V. (2011). An Exploration of Applying System Criminality to Arms Trafficking. International Criminal Justice Review, 22–38.

Stohl, R. (2004). The Tangled Web of Illicit Arms Trafficking. Center for American Progress, 21-26. The United Nations Office on Drugs and Crime. (2019). The Illicit Market in Firearms. Vienna: The Doha Decleration. United Nations Office for Disarament Affairs. (n.d.). ATT Implementation Toolkit: Overview of Obligations. United Nations.

United Nations Office for Disarmament Affairs. (n.d.). Small arms: Trade & brokering. Retrieved from United Nations: https://www.un.org/disarmament/convarms/trade-brokering/