Project 1 Report

Alexandre Lockhart

10/18/2020

Overall objectives

The war of correlates projects curates war data from 1814 globally through 2014 from all around the world. Variables such as deaths, outcome status, perceived initiator, periods of exposure, forces involved, whether the war was internationalized and spread and many other variables have been collected. Additional datasets such as economic factors, etc. have also been collected and tabulated with multiple datasets but the previously mentioned variables were of main interest for this project.

Due to the time period of curation (1814-2014), and the onset of the Monroe Doctrine in 1823, the central interest in this project was to look at the role of Americas (or the US and South America) versus the rest of the world in initiator-defined outcomes. This was historic in that in set a role of the United States in dictating colonial policy within the hemisphere up as north of Canada through the southern tip of Latin America in order to mandate control and keep international powers at bay cementing a power structure that has lasted until modern 21st century. The word initiator is used a lot in this project and is defined as the perceived initial war aggressor. Recipient is the perceived country who is the recipient of the initiator's attack.

Initially the project involved looking at initial descriptive relationships of the Americas versus non-Americas conflict via tables and graphs in the domestic wars only dataset. Variables such as deaths, type of conflict, internationalization of conflict, the Americas indicator of interest, exposure time of given conflict, and initiator created variables such as the absolute difference in initiator versus recipient deaths, and relative difference in the initiator to recipient deaths were visually examined.

Aim 1 of the project involved looking at Americas versus non-Americas in initiator determined outcomes such as: absolute difference in initiator versus recipient deaths and relative difference in initiator versus recipient deaths, . Initial basic glm models would be assessed while including war type, start year, and number of days of exposure. This would be repeated for the absolute and relative difference outcomes and then repeated using a training and testing set to assess prediction. The sensitivity of prediction would be assessed via imputation of deaths where missing, recomputing absolute and relative initiator variables of interest, and re-doing the prediction models.. The entire process would be repeated on an internationalized dataset, or one which adds forces available to those in conflict as well as adds additional wars perceived to be linked to the initial model set on an international scale.

Aim 2 was multi-fold: to cluster wartype, the Americas indicator variable, time since war started since 1814 basically, whether the war was internationalized, duration of exposure (days), initiator deaths, recipient deaths, and relative difference in deaths via a distance metric for mixed data types. Besides cluster assessment and performance evaluation, the clusters were then descriptively assessed with clustering variables as well as the conflict outcome for pattern assessment.

The final aim 3 was to look at network community detection based on weighting relative difference in deaths. A network data structure was made utilizing these weights. The goal was to descriptively assign membership and look at potentially separating attributes for a given community membership.

Preprocessing

Type of war was consolidated to Civil War: Central Control, Civil War: Local issues, and then other to account for sparsity of the last two categories. It is important to know that an internationalized dataset is also used (not shown) which adds wars to relevant wars that had conflict extend outside their local boundaries and also included forces available for each side in the conflict.

Initiator/recipient deaths and then the final outcome ('outcome E') was created by an indicator of initiator in the dataset as well as variables for side A/side B deaths and pattern matching in their construction. Imputations of the initiator and recipient death variables were created to account for missingness and eventual sensitivity analyses in prediction.

Test.An initial table 1 is shown below of demographics by the Americas and non-Americas indicator variable of main interest in the project. Not surprisingly, the number of regions and countries, etc. are going to have more wars representing 5/7 countries in the 'Not in Americas' category. Among the Americas, the distribution of wars classified as central control dominated proportionally (81%) while the war types were more balances in the not in Americas group. While the variability was high, the mean total days of war in the Americas versus Not in Americas was over 124 days less. Absolute difference in deaths (Initiator-recipient deaths) showed a much higher liability towards the initiator among conflicts in the Americas relative to Not in Americas. The time since curation (or time since initial curation of all wars (1814)) was much further away for conflicts Not in the Americas indicating (at least from a curation point of view) a lot more wars being represented in the latter half of the 20th century in comparison to wars in the Americas.The outcomeC variable is just to give a raw display of the outcome categories while outcomeE collapses and figures out the initiator/recipient deaths. In this case, the recipient appears to, on average 'win' more wars in the Americas based conflicts relative to the not in Americas group.

```
library(rmarkdown)
library(png)
library(compareGroups)

T1=readRDS('/home/rstudio/Overall_plots/plot_files/Table1.rds')

createTable(T1,show.p.overall=FALSE)
```

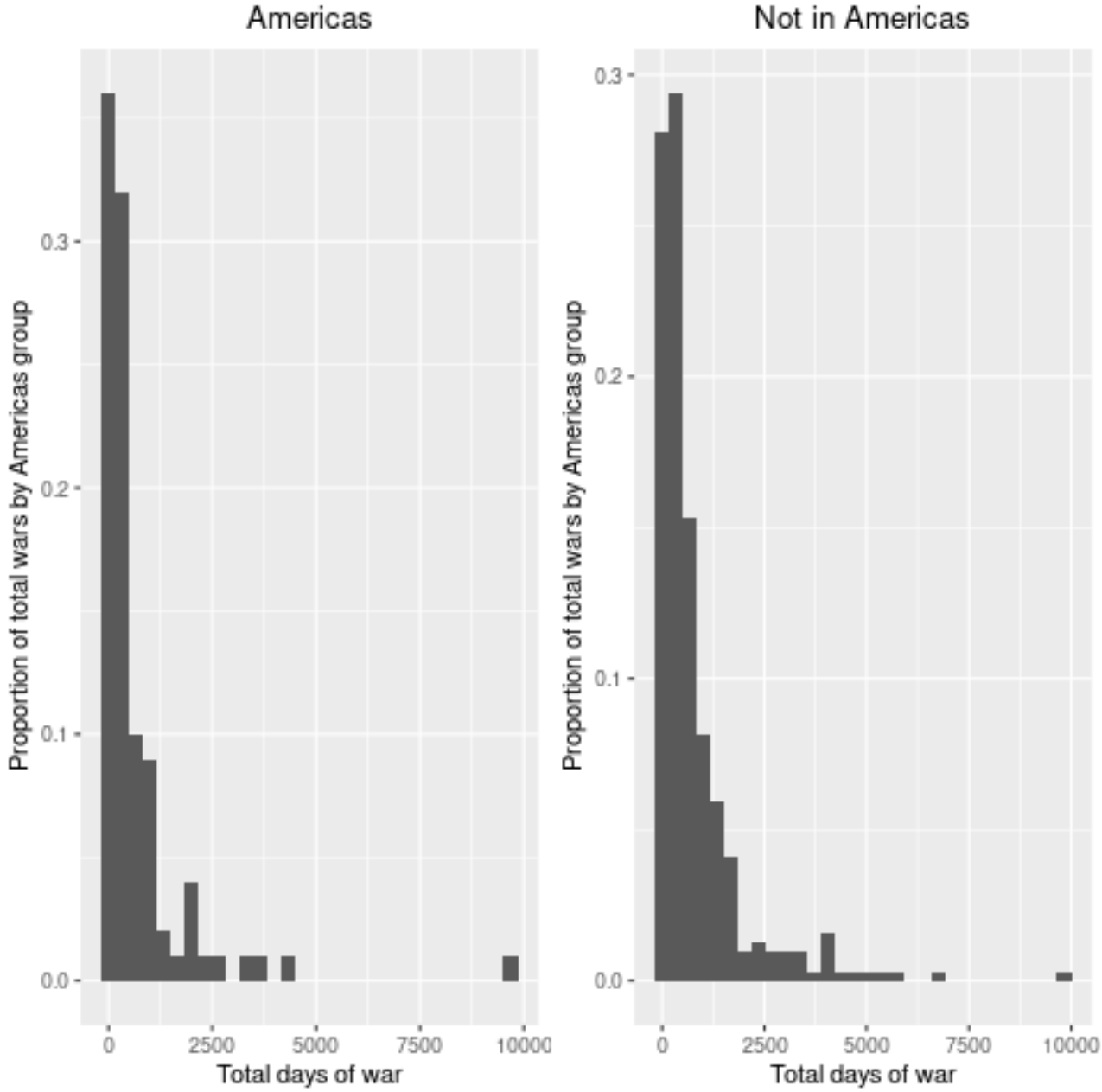
```
##
## --------Summary descriptives table by 'Americas'---------
##
## ____________________________________________________________________________
##                                          Americas     Not in Americas
##                                           N=100            N=320
## ----------------------------------------------------------------------------
## Type of War:
##     Civil War: Central Control             81 (81.0%)      142 (44.4%)
##     Civil War: Local Issues                17 (17.0%)      138 (43.1%)
##     Intercommunal                          0 (0.00%)        28 (8.75%)
##     Regional Internal                      2 (2.00%)        12 (3.75%)
## Total days of war                          658 (1215)       782 (1153)
## Initiator Deaths                          6473 (27670)     8084 (41390)
## Recipient Deaths                          8161 (38465)     9016 (54675)
## Relative difference in initiator deaths   0.01 (0.02)      0.01 (0.04)
## Absolute difference in initiator deaths -2767.28 (14965) -1723.58 (30636)
## Time since first curation  (years)         76.7 (44.0)       121 (58.3)
## Total deaths                              17013 (66326)    23452 (96210)
## Internationalized                          0.05 (0.22)      0.28 (0.45)
## OutcomeE:
##     Initiator Won                          10 (10.0%)       44 (13.8%)
```

```
##     Other                                61 (61.0%)      248 (77.5%)
##     Recipient Won                        29 (29.0%)       28 (8.75%)
## ----------------------------------------------------------------------
```

Initial plots

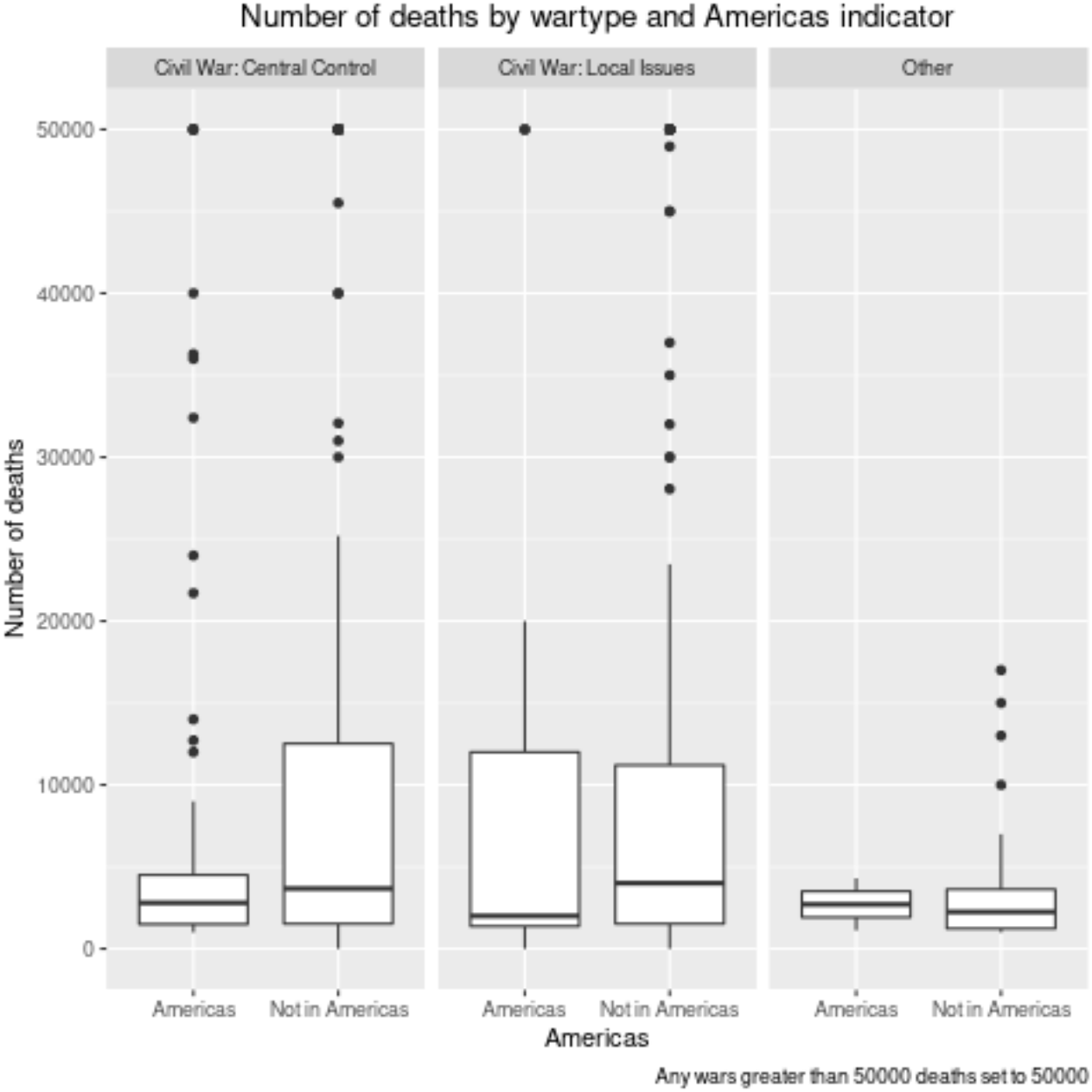
Exposure time on study

The below plots show the proportion of total wars that originated in Americas and Not in the Americas. Not surprisingly the total N (for both plots) is much greater (table 1 above) for the not in Americas wars, the distributions of war exposure and duration are roughly very similar in the past 200 years.



Battle deaths by the Americas

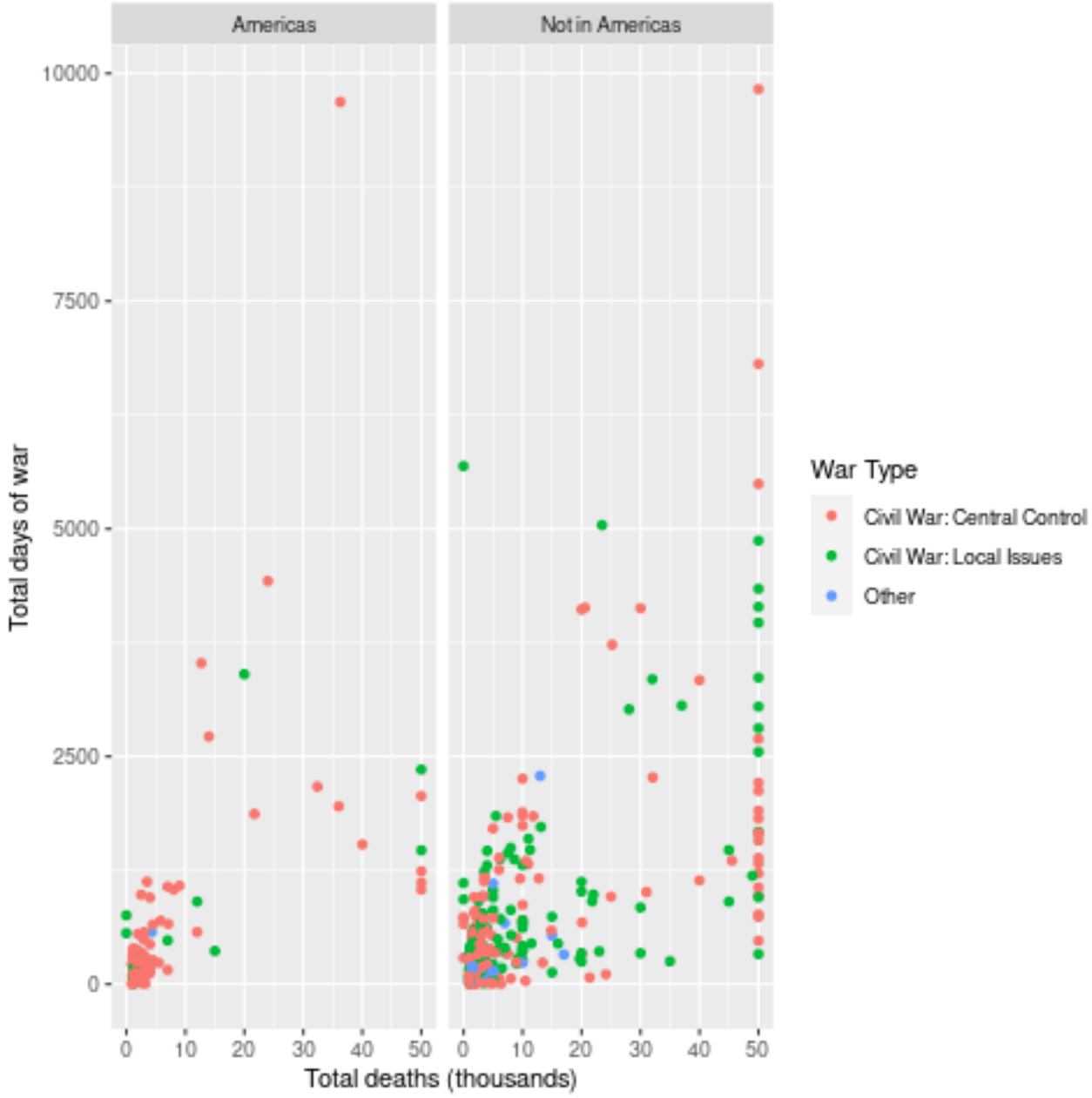
The below shows the distribution of deaths in the number of deaths by war type and Americas status. For the most part they look balanced. It should be noted that an upper limit of 50000 deaths was created in order to account for plot interpretability and to account for several wars with total deaths far greater than 50000. In the Americas group there appears to be much higher variability in the number of deaths for wars over local issues in comparison to wars over central control.



War type deaths Americas

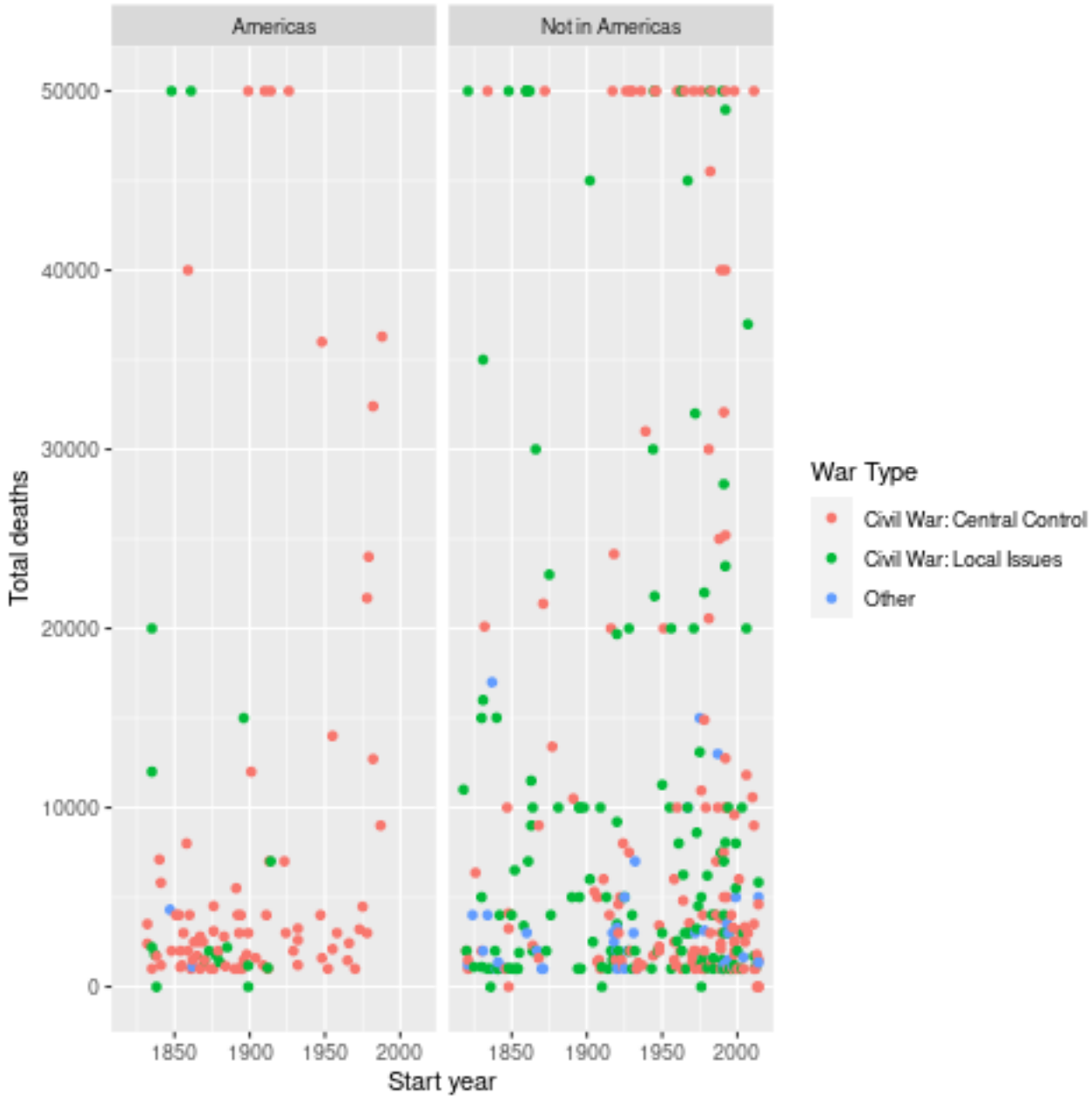
Adding the component of total war exposure it was also not surprising to see large proportion of wars with lower deaths having lower exposure time. For war type, no particular conflict type stood-out but proportionally, the not in Americas group has a larger variability in exposure time/deaths than wars in the

the Americas which tended to cluster more on average in the lower quadrant.



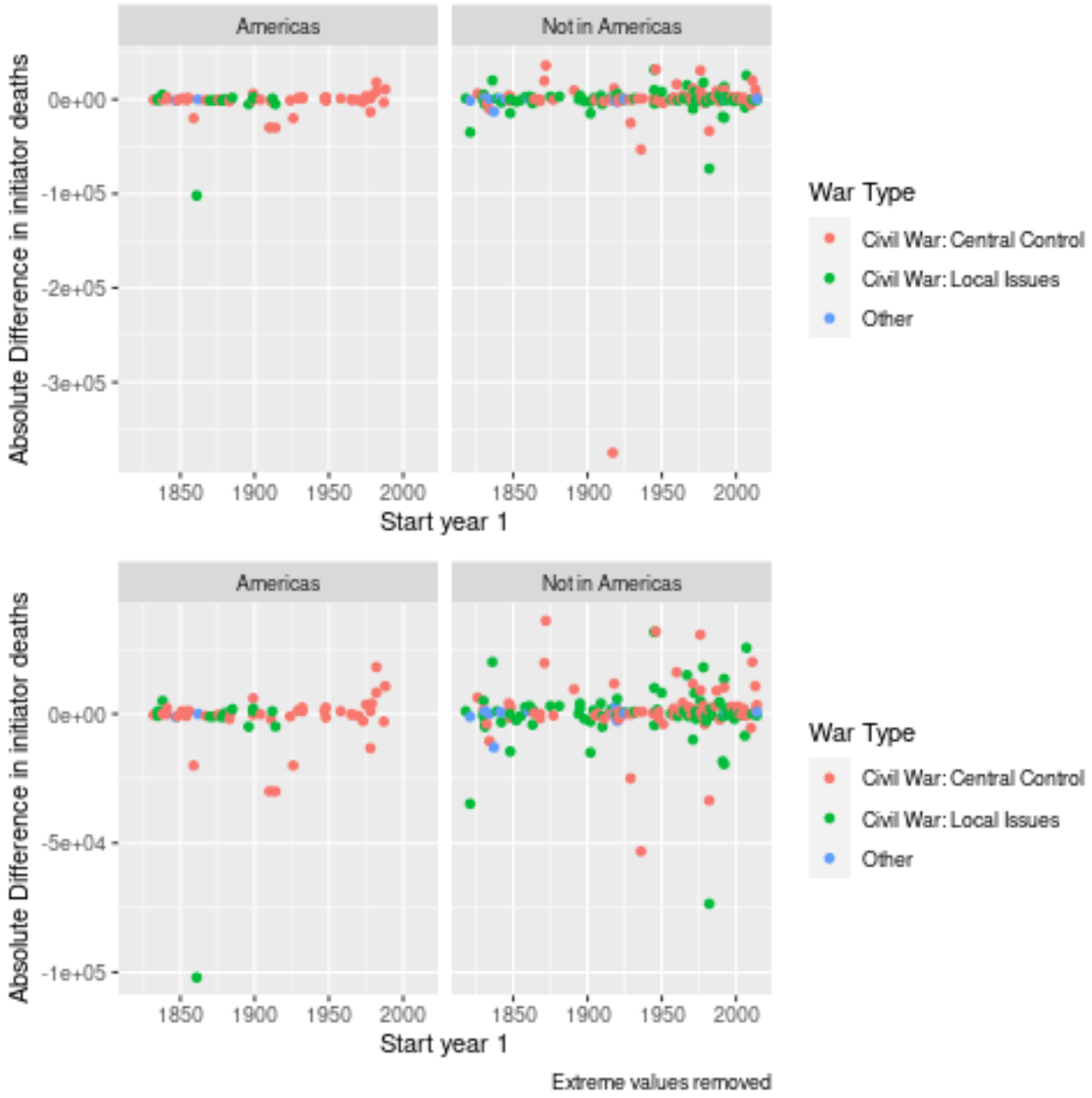
Start year

A plot looking at deaths based on war start year. From 1900s onwards the civil war for central control seemed to be most prevalent in the Americas while diversity of war type was prevalent across the 200 year period in the non-Americas group. As mentioned in the time to curation in table 1, the not in americas conflicts appear to have a considerable more set of curated conflicts in the latter half of the 20th century relative to conflicts in the Americas. Maybe one opinion is the military hegemony in this hemisphere by the United States to sort of dominate a lot of conflict for this hemisphere post WW-II in comparison to the rest of the world. There does not appear to be a linearly increasing or decreasing trend in deaths based on time of war and war conflict in either group.



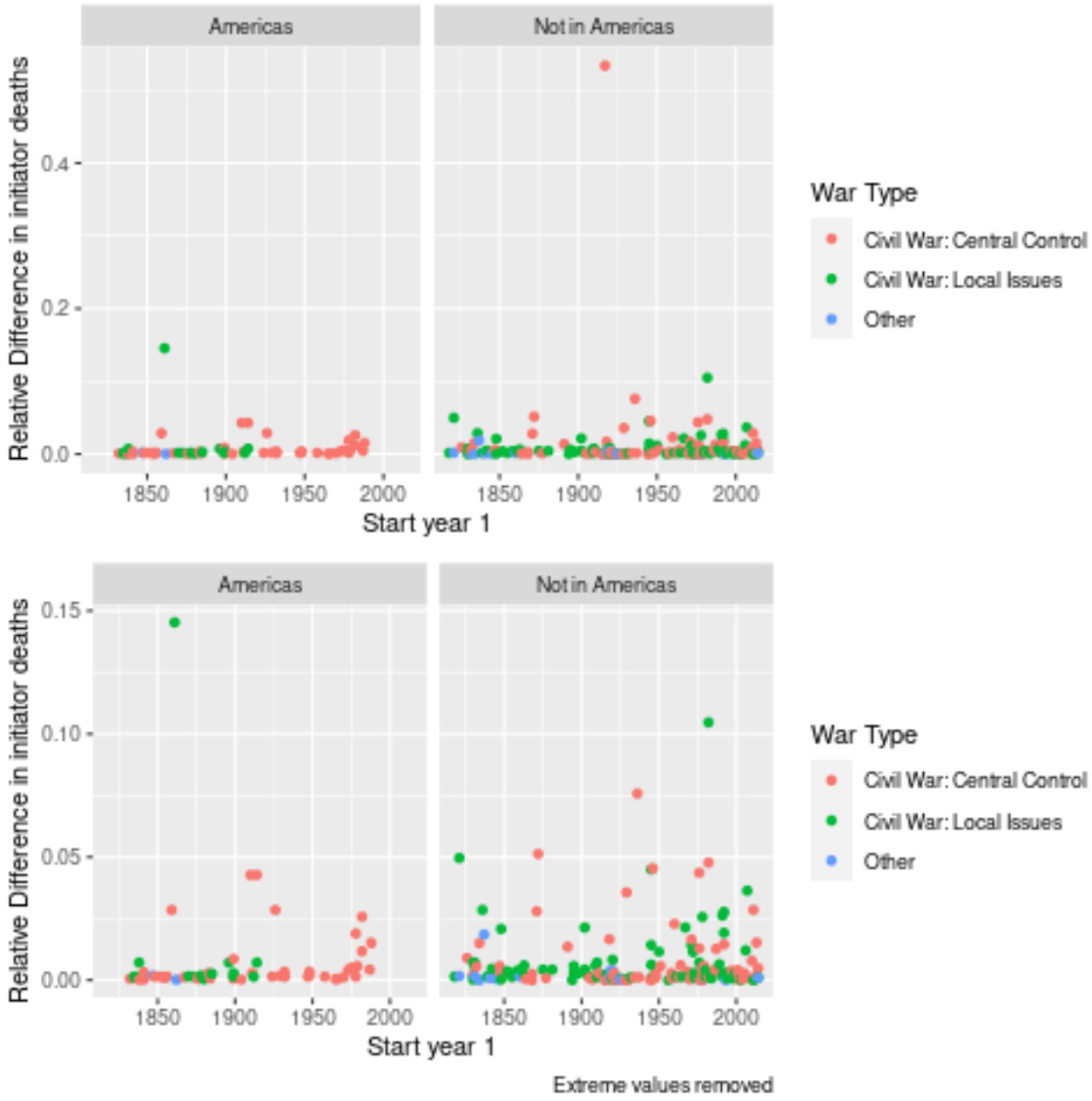
Absolute difference in deaths

Defined as the initiator deaths-recipient deaths, once several influential points (not outliers) were removed one can see a large proportion of initiator deaths more in the negatives from the 1900s onwards in the not in Americas group. Maybe this relates to some sort of potential increased ability in war recipients being able to foresee or address conflict. Maybe it could be greater familiarity with a 'home' area in response somehow.



Relative difference in deaths

This metric was more complicated defined as:abs(InitiatorDeaths-RecipientDeaths)/max(abs(RecipientDeaths),abs(InitiatorD the initiator deaths-recipient deaths, once several influential points (not outliers) were removed visually show a larger proportion of deaths among the initiators from the 1900s onwards.



Domestic simple model

Generalized linear models showing modeled the outcome of absolute difference in deaths and then relative difference in deaths by the Americas variable, starting year, days of exposure, and war type. In both model sets only an association was seen with days of exposure and relative increase in initiator deaths.

```
##                                 Description                     Variables
## 1 Absolute difference in deaths non-imputed                   (Intercept)
## 2                                                 AmericasNot in Americas
## 3                                                            StartYR_Norm
## 4                                                              WDuratDays
## 5                                             WarTypeDCivil War: Local Issues
## 6                                                            WarTypeDOther
```

```
##   Estimate Std err p-val
## 1 -5418.86 4788.97  0.26
## 2  -301.57 4421.93  0.95
## 3    28.39      34  0.40
## 4    -0.26    1.69  0.88
## 5  1414.91  3918.2  0.72
## 6  2748.99 7400.78  0.71
```

```
##                                 Description                      Variables
## 1 Relative difference in deaths non-imputed                    (Intercept)
## 2                                                    AmericasNot in Americas
## 3                                                              StartYR_Norm
## 4                                                                WDuratDays
## 5                                                 WarTypeDCivil War: Local Issues
## 6                                                                WarTypeDOther
##   Estimate Std err  p-val
## 1    0.008   0.006 0.1996
## 2    0.006   0.006 0.3192
## 3        0       0 0.2934
## 4        0       0 0.0048
## 5   -0.005   0.005 0.3882
## 6   -0.009    0.01 0.3520
```

Domestic prediction non-imputed model

A training and testing set (evenly split) using 5 fold cross-validation was done. Generalized linear models showing modeled the outcome of absolute difference in deaths and then relative difference in deaths by the Americas variable, starting year, days of exposure, and war type. In both model sets only the Americas versus non-Americas showed an association in the number of initiator deaths relative to the recipient. The prediction in the given test set, however, R2=.008, was very low. Also an association was seen with days of exposure and relative increase in initiator deaths.

```
##                                 Description     R2
## 1 Absolute difference in deaths non-imputed 0.0093
## 2
## 3
## 4
## 5
## 6
```

```
##                           Variables Estimate Std err p-val
## 1                       (Intercept) -4474.31 2944.33 0.131
## 2         'AmericasNot in Americas'  5415.25 2643.04 0.042
## 3                       StartYR_Norm    17.86   19.79 0.368
## 4                         WDuratDays    -0.58    1.11 0.600
## 5 'WarTypeDCivil War: Local Issues' -3528.15 2289.06 0.125
## 6                      WarTypeDOther -1293.59 4467.17 0.773
```

```
##                                 Description     R2
## 1 Relative difference in deaths non-imputed 0.0045
## 2
## 3
## 4
## 5
## 6
```

```
##                           Variables Estimate Std err  p-val
```

```
## 1                        (Intercept)   0.0045  0.0034   0.20
## 2         'AmericasNot in Americas'  -0.0029  0.0031   0.34
## 3                       StartYR_Norm -1.7e-05 2.3e-05   0.47
## 4                         WDuratDays  8.1e-06 1.3e-06 <0.001
## 5 'WarTypeDCivil War: Local Issues'   0.0036  0.0027   0.18
## 6                       WarTypeDOther  -0.0019  0.0052   0.71
```

Domestic prediction imputed models

The common theme across models was prediction was very low between training and test sets. An association, however, was seen in days and Americas was very close in the relative difference in initiator deaths. Date was re-checked and while certainty of curation seems reasonable (some wars simply were relatively more brutal than others), a prediction check of the imputed deaths while removing the most extreme war (375000 death difference), showed a slight gain in prediction (not shown but R2=.013 in comparison to .002)

```
##                             Description      R2
## 1 Absolute difference in deaths imputed 9.7e-06
## 2
## 3
## 4
## 5
## 6
##                            Variables Estimate Std err p-val
## 1                        (Intercept) -1486.45 2208.78  0.50
## 2         'AmericasNot in Americas'  1630.13  2287.7  0.48
## 3                       StartYR_Norm     0.14   15.98  0.99
## 4                         WDuratDays     0.87    0.69  0.21
## 5 'WarTypeDCivil War: Local Issues' -1203.73 1996.03  0.55
## 6                       WarTypeDOther -3374.92 2992.47  0.26
```

```
##                             Description     R2
## 1 Relative difference in deaths imputed 0.0044
## 2
## 3
## 4
## 5
## 6
##                            Variables Estimate Std err  p-val
## 1                        (Intercept)  0.00068  0.0026   0.79
## 2         'AmericasNot in Americas'   0.0018  0.0027   0.50
## 3                       StartYR_Norm  2.7e-05 1.9e-05   0.14
## 4                         WDuratDays  3.5e-06   8e-07 <0.001
## 5 'WarTypeDCivil War: Local Issues'   0.0026  0.0023   0.26
## 6                       WarTypeDOther   0.0048  0.0035   0.16
```

Internationalized Non-imputed

In the internationalized models they took into account wars also linked to domestic wars and additional variables such as initiator forces and recipient forces were taken into account in models. While predictions remained weak, days of exposure popped out in models and recipient forces showed an association in relative difference in association models.

```
##                                      Description    R2
## 1 Intl Absolute difference in deaths non-imputed 0.011
```

```
## 2
## 3
## 4
## 5
## 6
## 7
## 8
##                          Covariates Estimate Std err p-val
## 1                       (Intercept)  -2875.7 2354.04 0.224
## 2        'AmericasNot in Americas'  3177.52 2265.41 0.163
## 3                      StartYR_Norm    15.32   16.08 0.342
## 4                       WDuratDays     -2.2    0.98 0.027
## 5 'WarTypeDCivil War: Local Issues' -2670.35 1858.04 0.153
## 6                     WarTypeDOther -1991.32 3945.77 0.614
## 7                   InitiatorForces        0       0 0.856
## 8                   RecipientForces        0       0 0.067
```

```
##                                     Description    R2
## 1 Intl Relative difference in deaths non-imputed 0.033
## 2
## 3
## 4
## 5
## 6
## 7
## 8
##                          Covariates Estimate Std err  p-val
## 1                       (Intercept)   -0.011   0.016 0.4647
## 2        'AmericasNot in Americas'  0.00013   0.013 0.9915
## 3                      StartYR_Norm  3.1e-05 0.00014 0.8263
## 4                       WDuratDays  3.4e-05   1e-05 <0.001
## 5 'WarTypeDCivil War: Local Issues'  -0.0037   0.012 0.7478
## 6                     WarTypeDOther  0.00079   0.025 0.9744
## 7                   InitiatorForces  8.1e-09 1.9e-08 0.6738
## 8                   RecipientForces  9.1e-09 2.8e-09 0.0013
```

Internationalized Imputed

In both models the recipient forces and days of exposure also showed an interesting result but nothing with Americas.

```
##                                   Description    R2
## 1 Intl Absolute difference in deaths imputed 0.011
## 2
## 3
## 4
## 5
## 6
## 7
## 8
##                          Covariates Estimate Std err p-val
## 1                       (Intercept)  -2875.7 2354.04 0.224
## 2        'AmericasNot in Americas'  3177.52 2265.41 0.163
## 3                      StartYR_Norm    15.32   16.08 0.342
```

```
## 4                         WDuratDays     -2.2    0.98 0.027
## 5 'WarTypeDCivil War: Local Issues' -2670.35 1858.04 0.153
## 6                       WarTypeDOther -1991.32 3945.77 0.614
## 7                    InitiatorForces        0       0 0.856
## 8                    RecipientForces        0       0 0.067
```

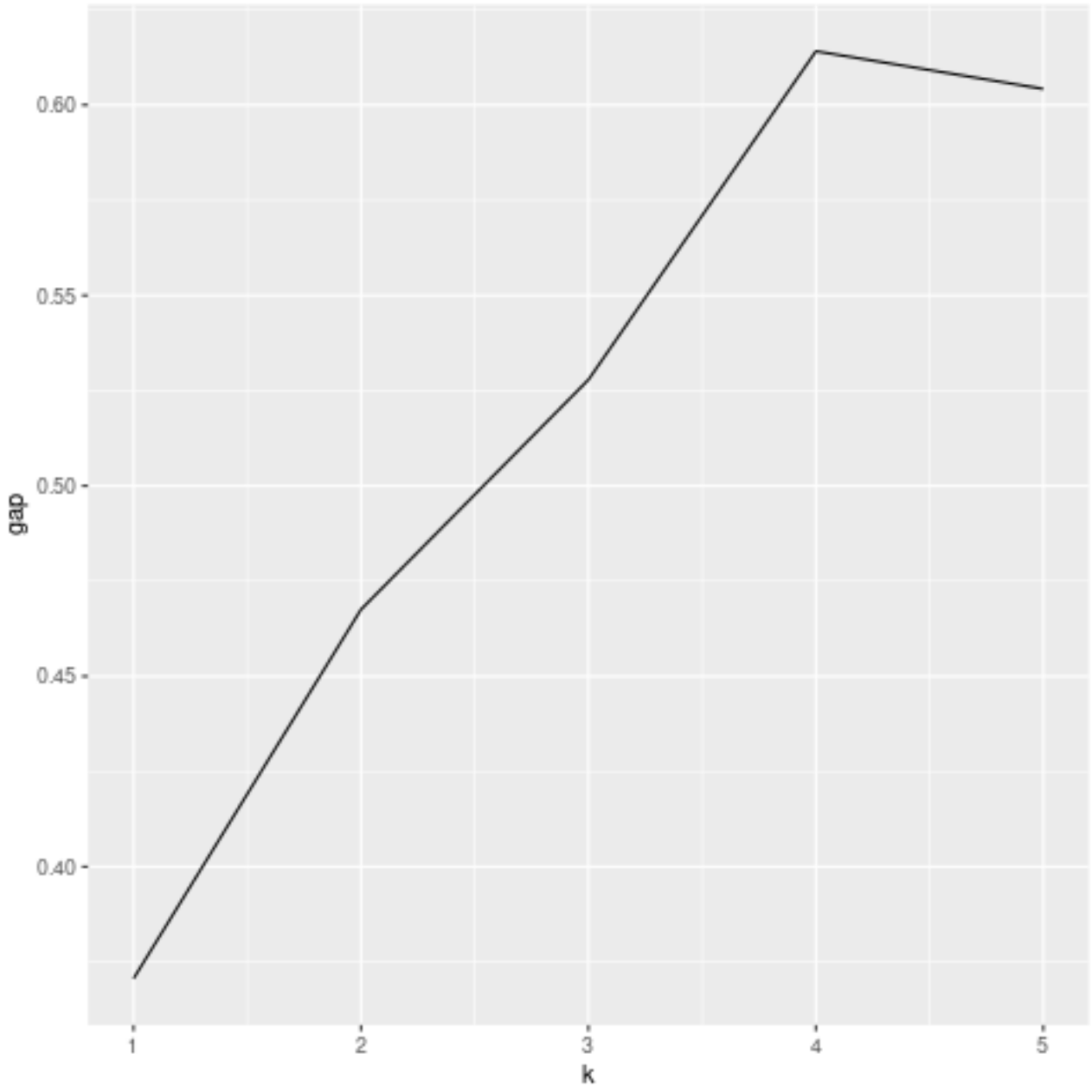
```
##                                 Description    R2
## 1 Intl Relative difference in deaths imputed 0.033
## 2
## 3
## 4
## 5
## 6
## 7
## 8
```

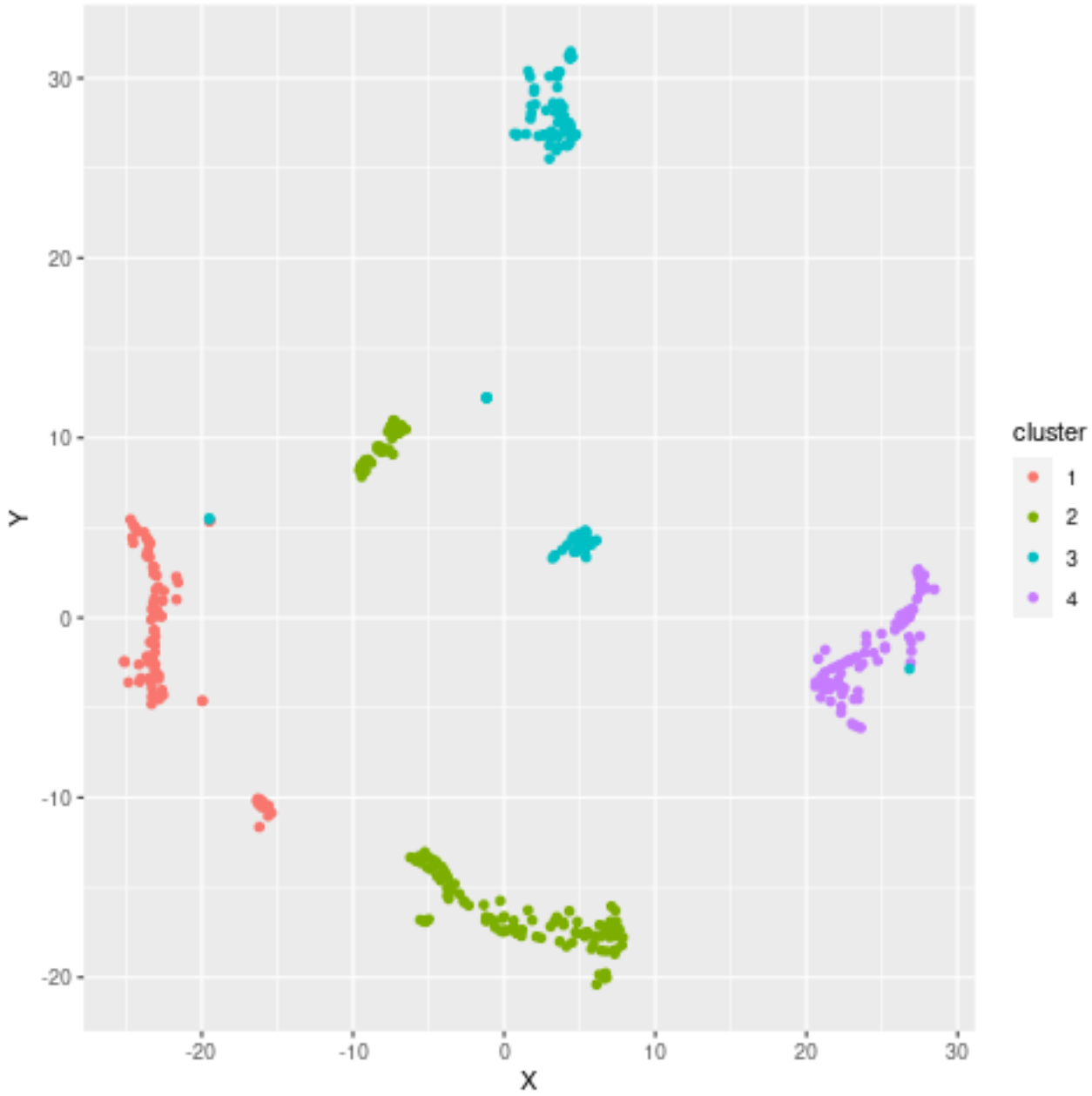
```
##                           Covariates Estimate Std err  p-val
## 1                         (Intercept)   -0.011   0.016 0.4647
## 2         'AmericasNot in Americas'  0.00013   0.013 0.9915
## 3                       StartYR_Norm  3.1e-05 0.00014 0.8263
## 4                         WDuratDays  3.4e-05   1e-05 <0.001
## 5 'WarTypeDCivil War: Local Issues'  -0.0037   0.012 0.7478
## 6                       WarTypeDOther  0.00079   0.025 0.9744
## 7                    InitiatorForces  8.1e-09 1.9e-08 0.6738
## 8                    RecipientForces  9.1e-09 2.8e-09 0.0013
```

Overall clustering

Aim 3 was multi-fold: to cluster wartype, the Americas indicator variable, time since war started since 1814 basically, whether the war was internationalized, duration of exposure (days), initiator deaths, recipient deaths, and relative difference in deaths. A distance metric for mixed data types (Gower) was used and clustering via KMeans. tSNE plots showed the segmentation of the variability of the gower's distance metric, and the clusters were visually plotted.

Below shows the top 2 TSNE dimensions by cluster and showing distinct separation throughout.


0.60
0.55
0.50
0.45
0.40
gap
1
2
3
4
5
k



Overall clustering by variables

As described above the four clusters showed a distance pattern across the variables in their construction. Clearly defined Americas (cluster 1), and non-Americas clusters were shown (2 through 4) which looked very interesting albeit a little fishy. The largest proportion of those where the recipient won was in the Americas cluster. Otherwise stalemates, compromises, etc. were the largest in all three of the specific clusters. Cluster 3 could be established as the highest difference in deaths for the initiator for some reason. Cluster 3 could also be seen to have the most modern (time since curation of 1814) conflict cluster with an average starting time of conflicts ranging from the late 1950s onwards.

```
##
## --------Summary descriptives table by 'cluster'---------
##
```

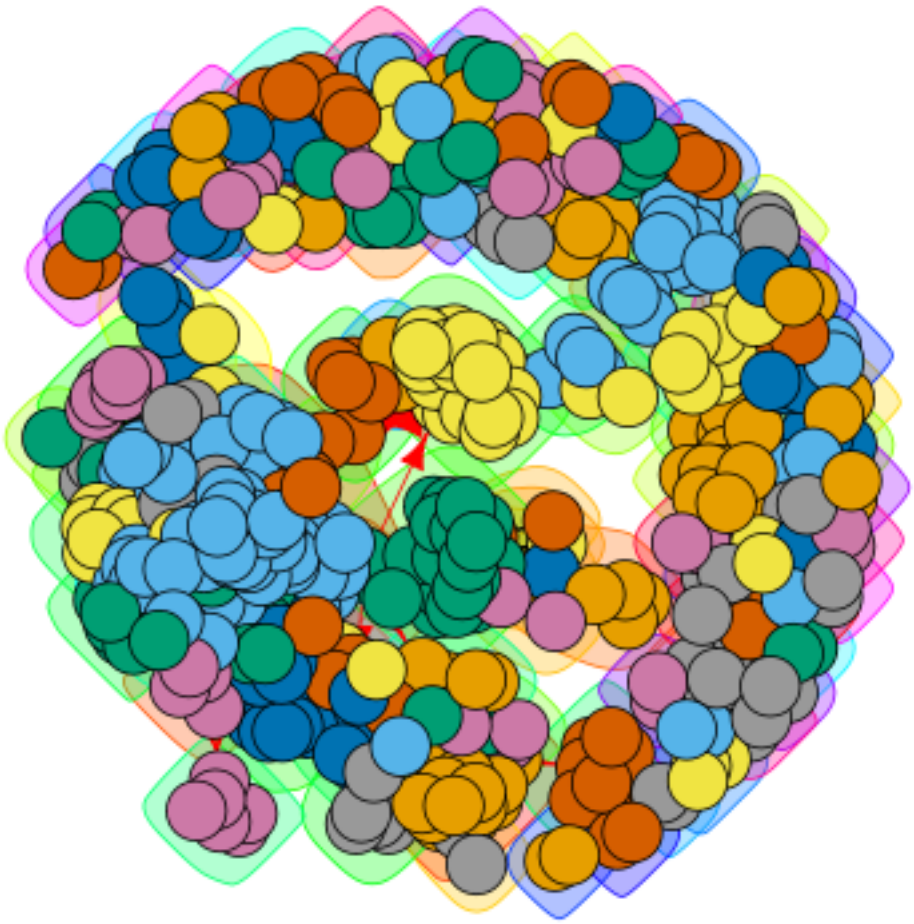
	1	2	3	4
	N=98	N=147	N=92	N=83
Americas:				
Americas	98 (100%)	0 (0.00%)	2 (2.17%)	0 (0.00%)
Not in Americas	0 (0.00%)	147 (100%)	90 (97.8%)	83 (100%)
OutcomeE:				
Initiator Won	10 (10.2%)	28 (19.0%)	4 (4.35%)	12 (14.5%)
Other	59 (60.2%)	115 (78.2%)	78 (84.8%)	57 (68.7%)
Recipient Won	29 (29.6%)	4 (2.72%)	10 (10.9%)	14 (16.9%)
WarTypeC:				
Civil War: Central Control	79 (80.6%)	0 (0.00%)	61 (66.3%)	83 (100%)
Civil War: Local Issues	17 (17.3%)	111 (75.5%)	27 (29.3%)	0 (0.00%)
Intercommunal	0 (0.00%)	26 (17.7%)	2 (2.17%)	0 (0.00%)
Regional Internal	2 (2.04%)	10 (6.80%)	2 (2.17%)	0 (0.00%)
WDuratDays	665 (1226)	674 (1041)	1073 (1450)	641 (880)
InitiatorDeaths	6589 (27942)	3267 (9287)	17903 (72554)	5558 (22675)
RecipientDeaths	8304 (38846)	3327 (13458)	21608 (97358)	4945 (23164)
RelDiffDeaths	0.01 (0.02)	0.01 (0.01)	0.02 (0.08)	0.01 (0.01)
AbsDiffDeaths	-2850.22 (15214)	-121.26 (10905)	-7100.94 (55926)	942 (8592)
StartYR_Norm	75.2 (43.3)	102 (60.5)	143 (50.3)	129 (51.5)
OutcomeC:				
Compromise	3 (3.06%)	16 (10.9%)	15 (16.3%)	5 (6.02%)
Conflict continues below war level	1 (1.02%)	13 (8.84%)	4 (4.35%)	7 (8.43%)
Side A wins	60 (61.2%)	77 (52.4%)	32 (34.8%)	43 (51.8%)
Side B wins	32 (32.7%)	14 (9.52%)	20 (21.7%)	19 (22.9%)
Stalemate	0 (0.00%)	14 (9.52%)	5 (5.43%)	4 (4.82%)
War ongoing as of end of 2014	1 (1.02%)	1 (0.68%)	7 (7.61%)	3 (3.61%)
War transformed into another War	1 (1.02%)	12 (8.16%)	9 (9.78%)	2 (2.41%)
TotalBDeaths	17320 (66971)	11898 (29275)	50105 (167912)	13853 (45569)
Intnl	0.03 (0.17)	0.00 (0.00)	0.99 (0.10)	0.00 (0.00)

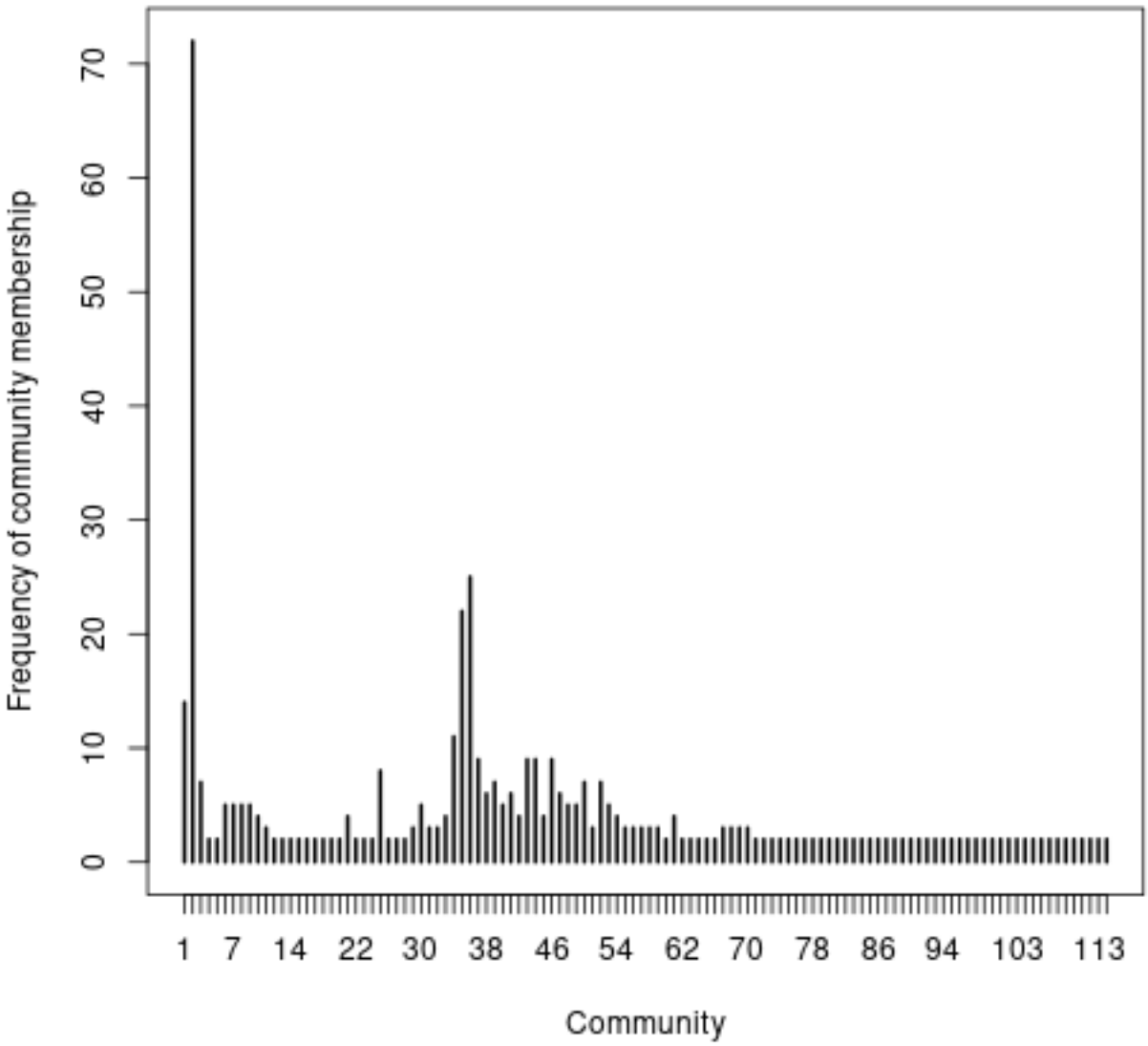
Aim 4 network composition

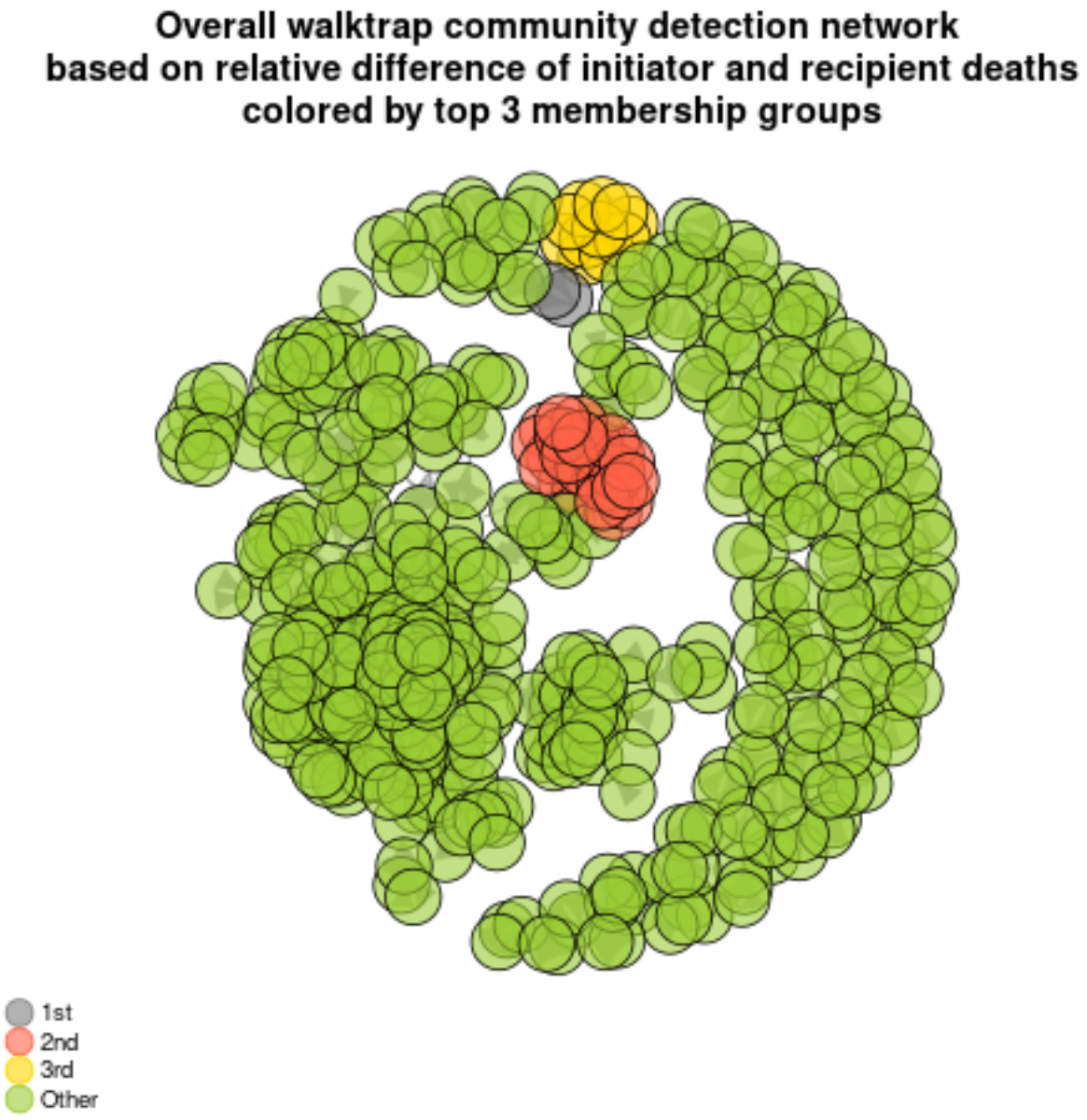
Several network plots were constructed. A walk trap community search algorithm was implemented and is generally a solid approach when not much is known about the given network structure. Weights between conflicts were normalized differences in conflicts and directed graph structure was created representing initiator to recipient relationships.

The first overall network plot shows all the wars together and their post memberships. Roughly a 100 communities were found, and rather than pruning, the second plot shows the distribution of community membership by wars representing the largest communities. The third plot shows the communities consolidated by the first three most prevalent communities and the fourth determined as 'other'. Edge attributes explaining these 3 particular communities and war characteristics which define them would be the next characteristic to find.

Overall walktrap community detection network
based on relative difference of initiator and recipient deaths

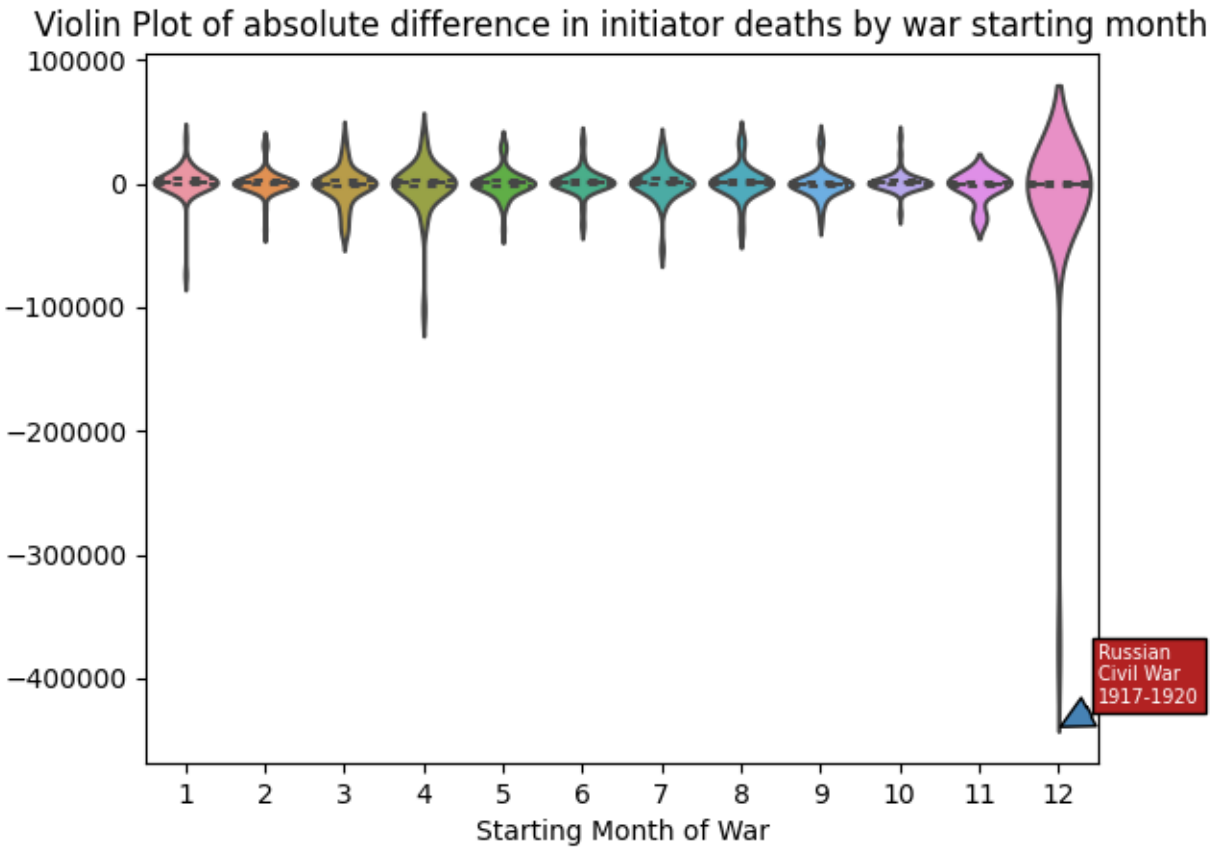



Frequency of community membership
0
10
20
30
40
50
60
70
1
7
14
22
30
38
46
54
62
70
78
86
94
103
113
Community



Python project plot

A violin plot of absolute difference in initiator deaths by calendar month was shown. The interest was in seeing if a general shape or variability by calendar month might be different for a given war start month. Not much was seen, however, though one war did standout in December, which was annotated (Russian Civil War 1917-1920). In this given war, the anti-Bolsheviks opposing Lenin's regime were perceived as the war 'initiators' despite fighting for independence and suffered the relatively highest number of initiator casualties relatively to the other side, of all the curated wars.



Conclusions

Descriptive plots showed some interesting patterns such as considerably less conflicts post WW-2 in the hemisphere occupied by North and South America relative to the rest of the world.

While the main effect was the Americas indicator variable, and these were technically confounders in the relationship (mainly days in almost all models and occasionally one of the forces available variables), a future consideration would be to actually use days of exposure as a modifier in the Americas/outcome relationship of all models, rather than as a confounder.

The most interesting part of the project turned out to be the initial clustering and descriptive representation with the key war variables and in particular the Americas variable. Prediction modeling initially wanted to look at sub groups or tree based models after the GLMs. But such low performance and variability in the differences in initiator and recipient deaths probably contributed to the poor performance. Rather than exploring the tree models, a mixed data clustering approach looked at the key war variables and generated some interesting results

Another initial aim to look at data-driven eras or model-based clustering as well as waves of a given war. Unfortunately the date curation for many waves beyond the first wave was more sparse than originally thought, and looking at time based eras did not have the same interest. Finally, the network community structure work found several clusters of communities but analysis is still in process as to glean interesting network properties from these communities. For instance, what makes thee three highest memberships unique among themselves and in comparison to conflicts which were in other communities? This is still in process. One difficulty in networks tends to be reproducibility so further methodology which takes this into

account will be assessed.