Project 1 Report

Alexandre Lockhart

10/4/2020

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.

Aim 1 of 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 2 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 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 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 4 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 An initial table 1 is shown below of demographics by the Americas and non-Americas indicator variable of main interest in the project. 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.

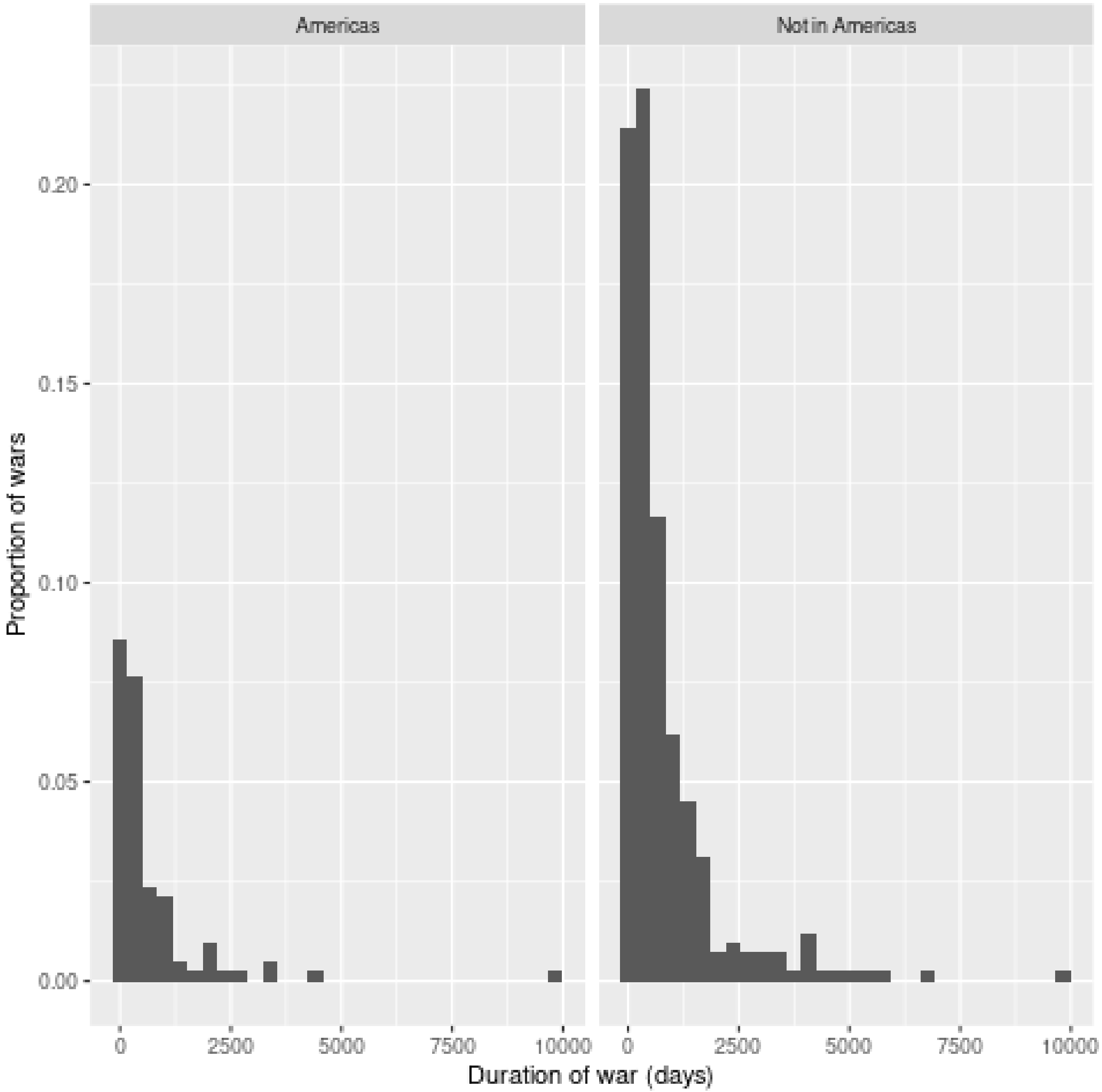
Initiator/recipient deaths and then the final outcome ('outcome E') was created by an indicator of initiator in the dataset as well as variables for side A/side B deaths and pattern matching in their construction. Imputations of the initiator and recipient death variables were created to account for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

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

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

createTable(T1)
```

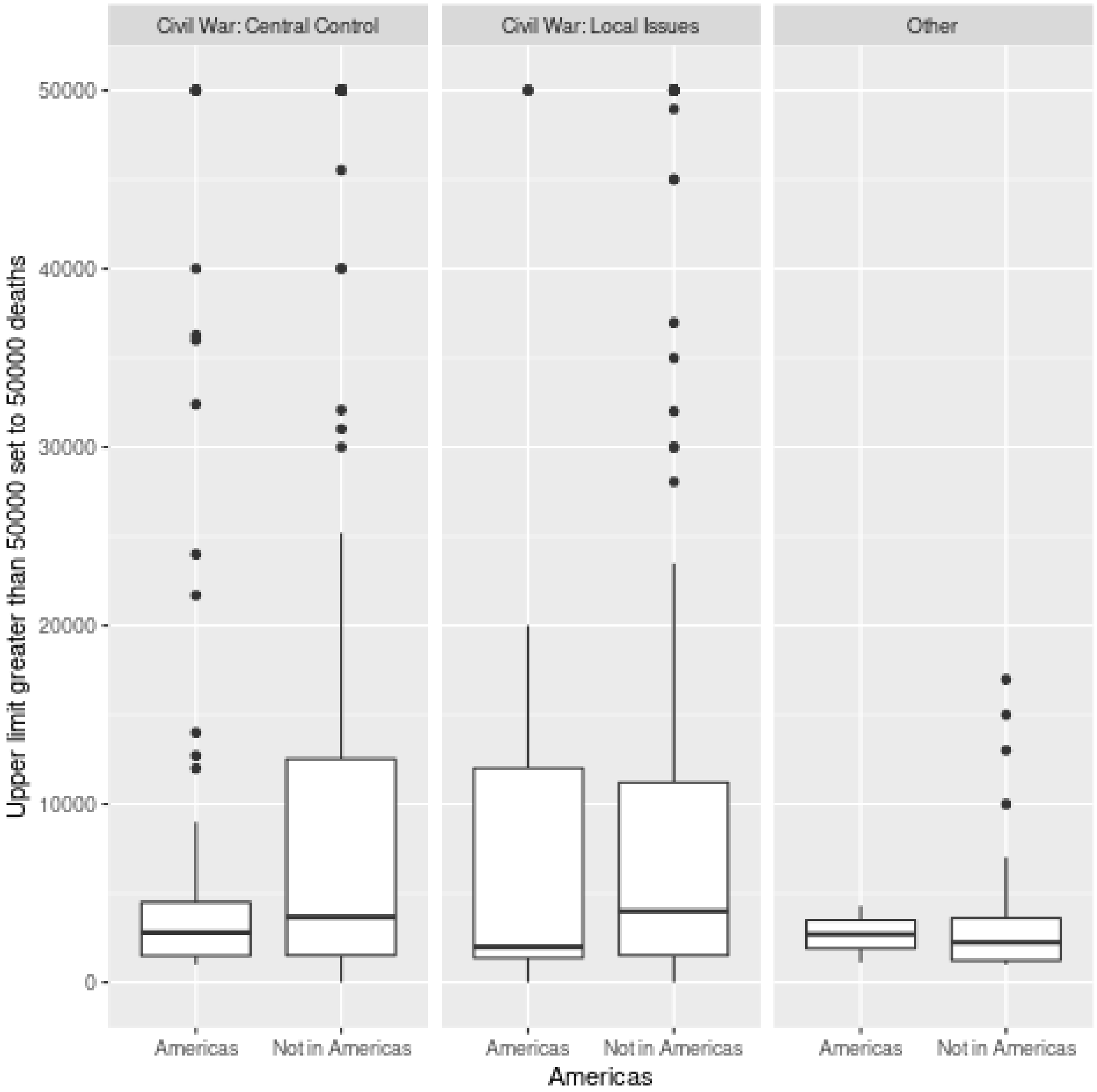
```
##
## --------Summary descriptives table by 'Americas'---------
##
## -----------------------------------------------------------------------------------
##                                             Americas     Not in Americas  p.overall
##                                              N=100            N=320
## -----------------------------------------------------------------------------------
## Type of War:                                                               <0.001
##     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)      0.368
## Initiator Deaths                          6473 (27670)    8084 (41390)     0.655
## Recipient Deaths                          8161 (38465)    9016 (54675)     0.862
## Relative difference in initiator deaths  -11.18 (76.7)    -20.22 (196)     0.613
## Absolute difference in initiator deaths -2767.28 (14965) -1723.58 (30636)   0.730
## Time since first curation  (years)        76.7 (44.0)      121 (58.3)     <0.001
## OutcomeC:                                                                     .
##     Compromise                             3 (3.00%)        36 (11.2%)
##     Conflict continues below war level     3 (3.00%)        22 (6.88%)
##     Side A wins                            60 (60.0%)      152 (47.5%)
##     Side B wins                            32 (32.0%)       53 (16.6%)
##     Stalemate                              0 (0.00%)        23 (7.19%)
##     War ongoing as of end of 2014          1 (1.00%)        11 (3.44%)
##     War transformed into another War       1 (1.00%)        23 (7.19%)
## Total deaths                             17013 (66326)    23452 (96210)    0.452
## Internationalized                         0.05 (0.22)      0.28 (0.45)    <0.001
## OutcomeE:                                                                  <0.001
##     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

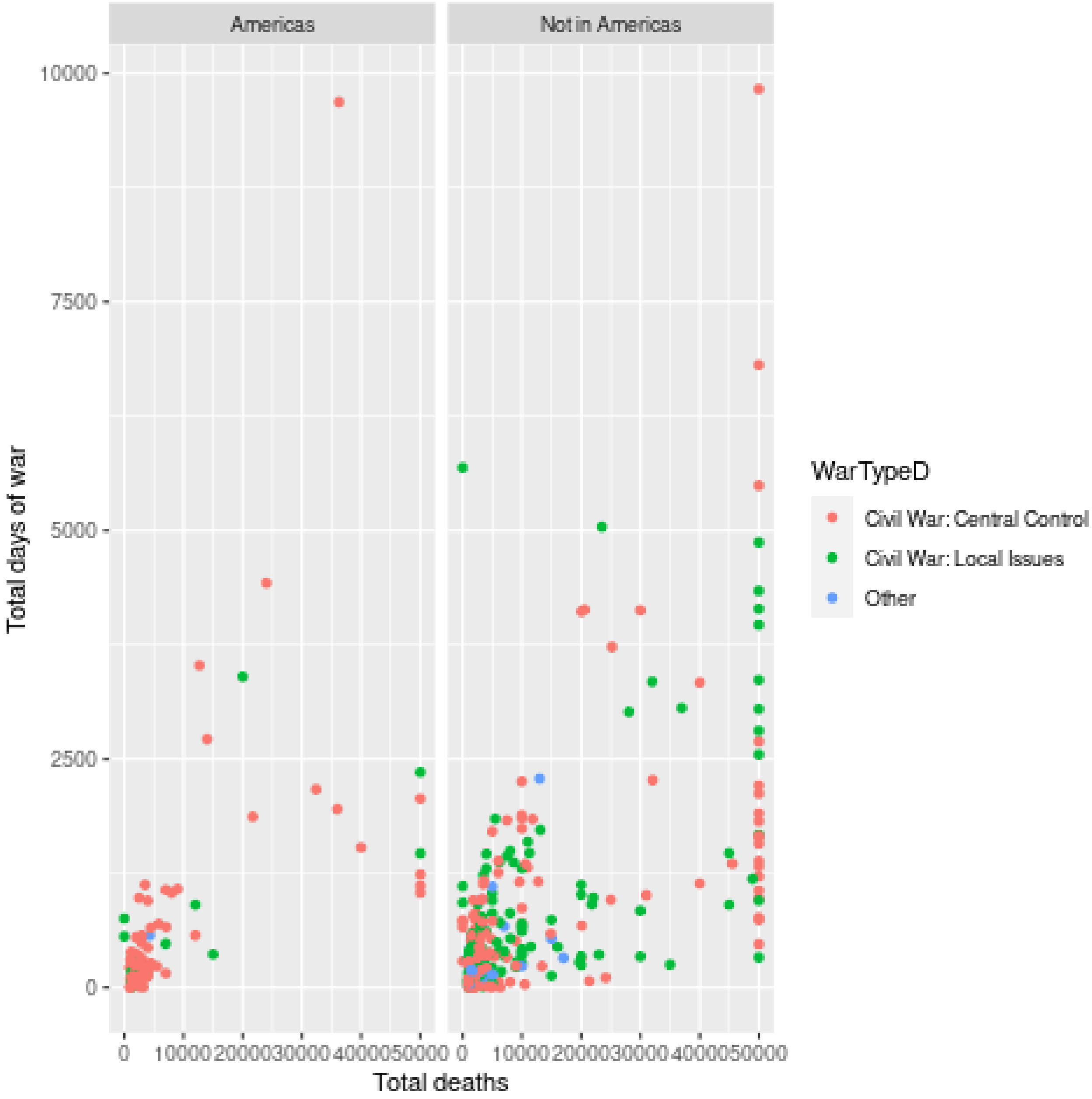
#Exposure time on study The below plots show the proportion of total wars that originated in Americas and Not in the Americas. Not surprisingly the total N (for both plots) shows an imbalance in distribution in the days of the wars. Accounting for this imbalance would be a future consideration in modeling but for now was left unaccounted.



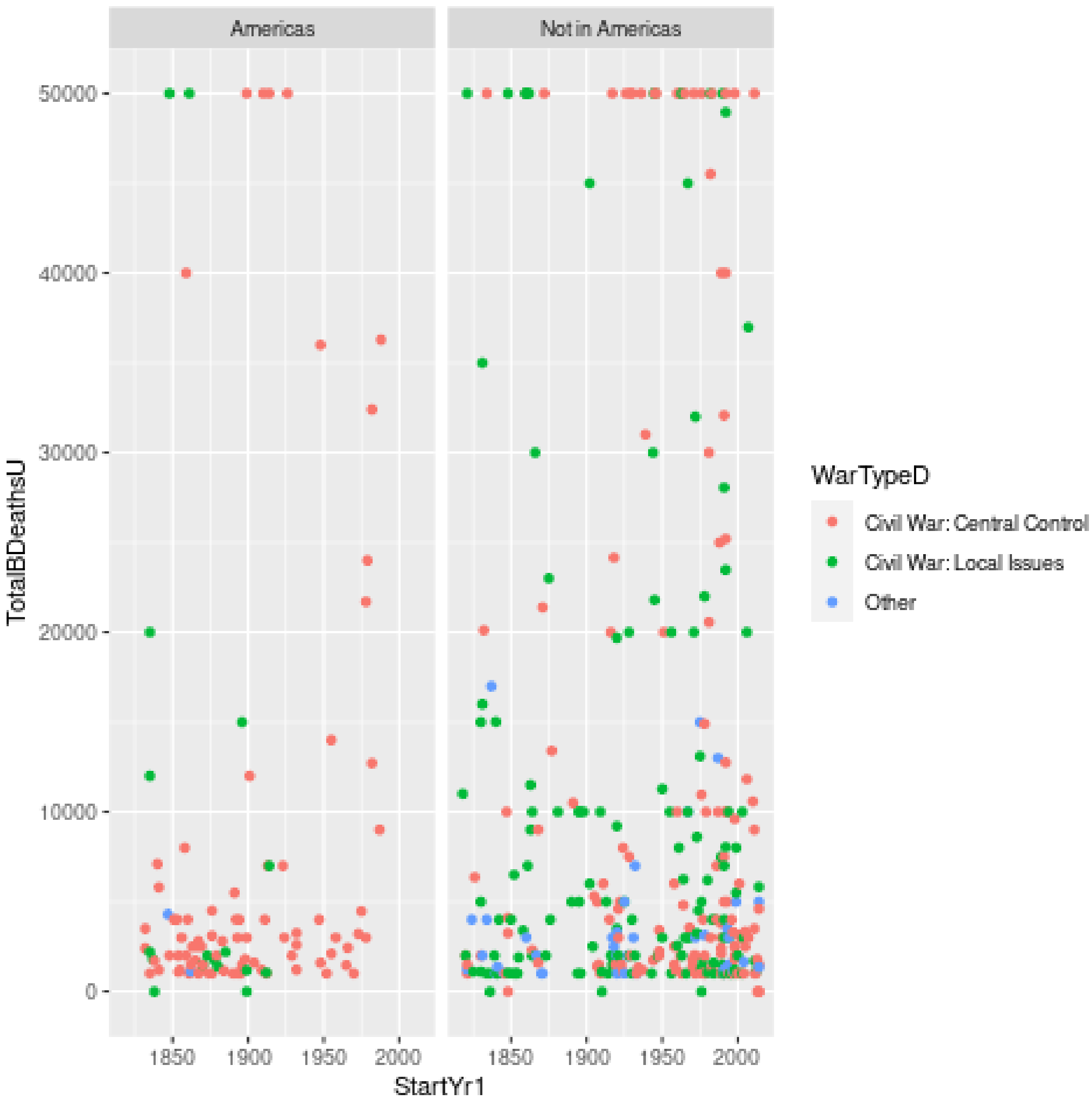
#Battle deaths by the Americas The below shows the distribution of deaths in the number of deaths by war type and Americas status. For the most part they look balanced. It should be noted that an upper limit of 50000 deaths was created in order to account for plot interpretability and to account for several wars with total deaths far greater than 50000.



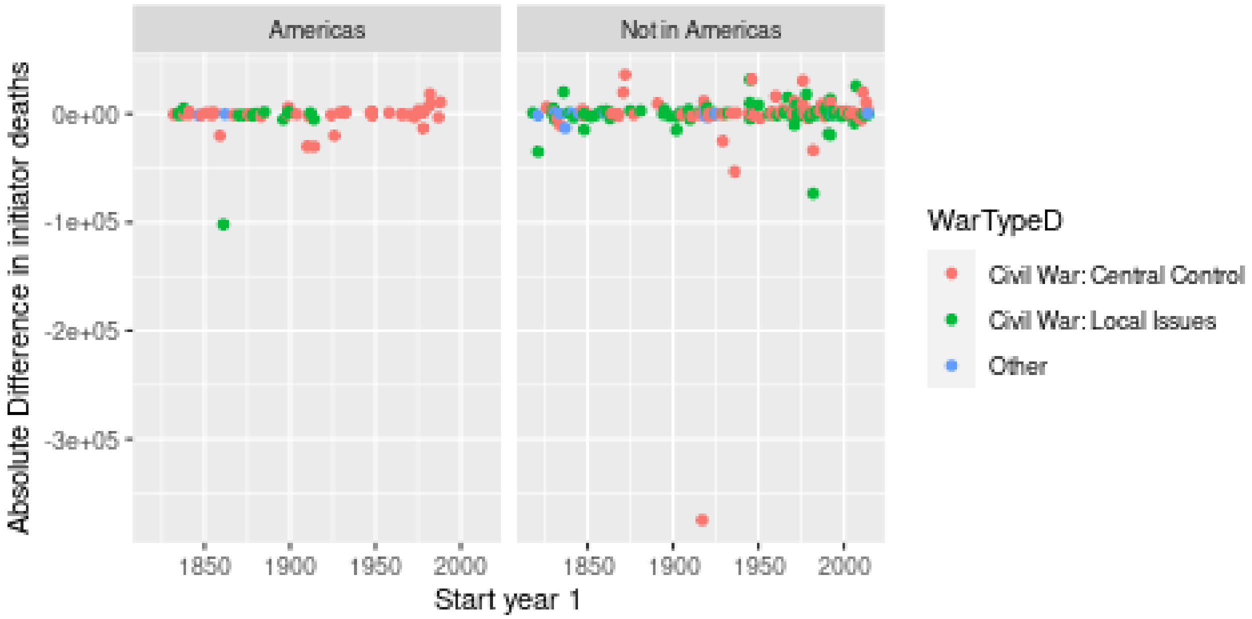
#War type deaths Americas Adding the component of total war exposure it was also not surprising to see large proportion of wars with lower deaths having lower exposure time. For war type, no particular conflict type stood-out.

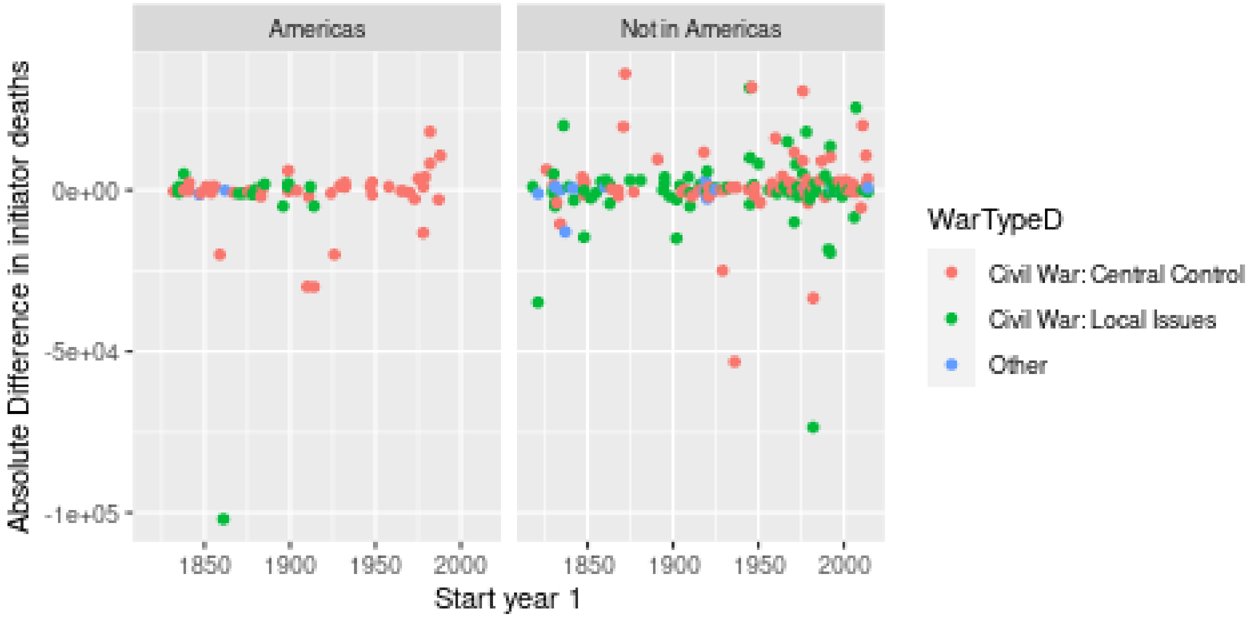


#Start year A plot looking at deaths based on war start year. From 1900s onwards the civil war for central control seemed to be most prevalent in the Americas while diversity of war type was prevalent across the 200 year period in the non-Americas group. There does not appear to be a linearly increasing or decreasing trend in deaths based on time of war and war conflict in either group.

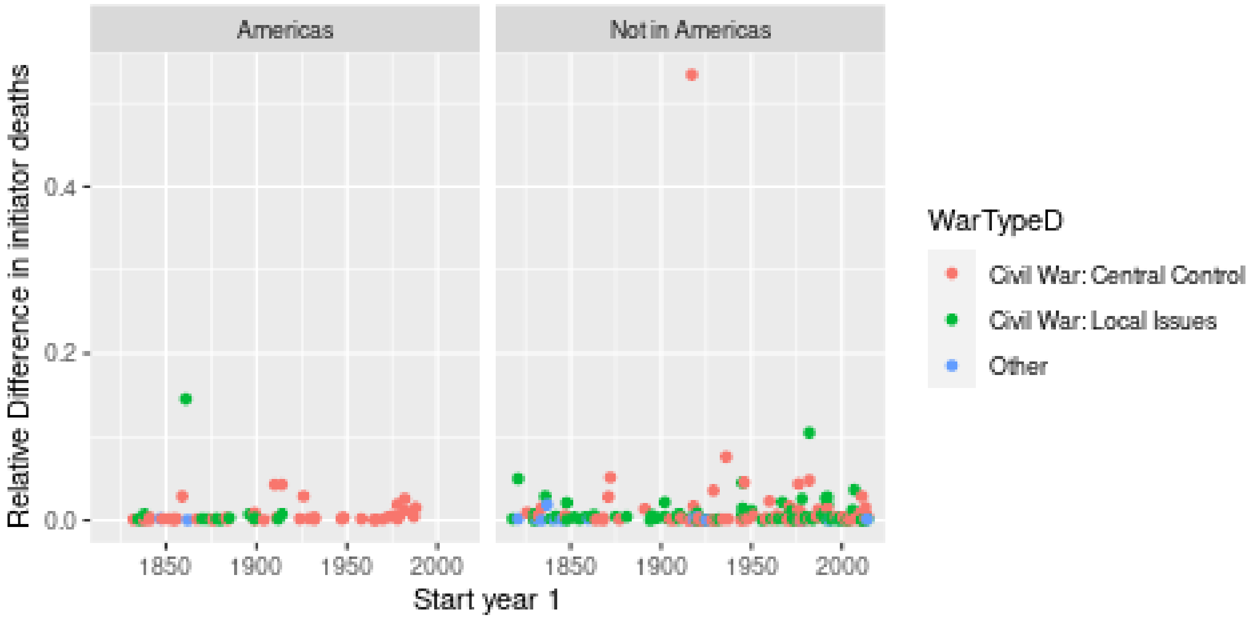


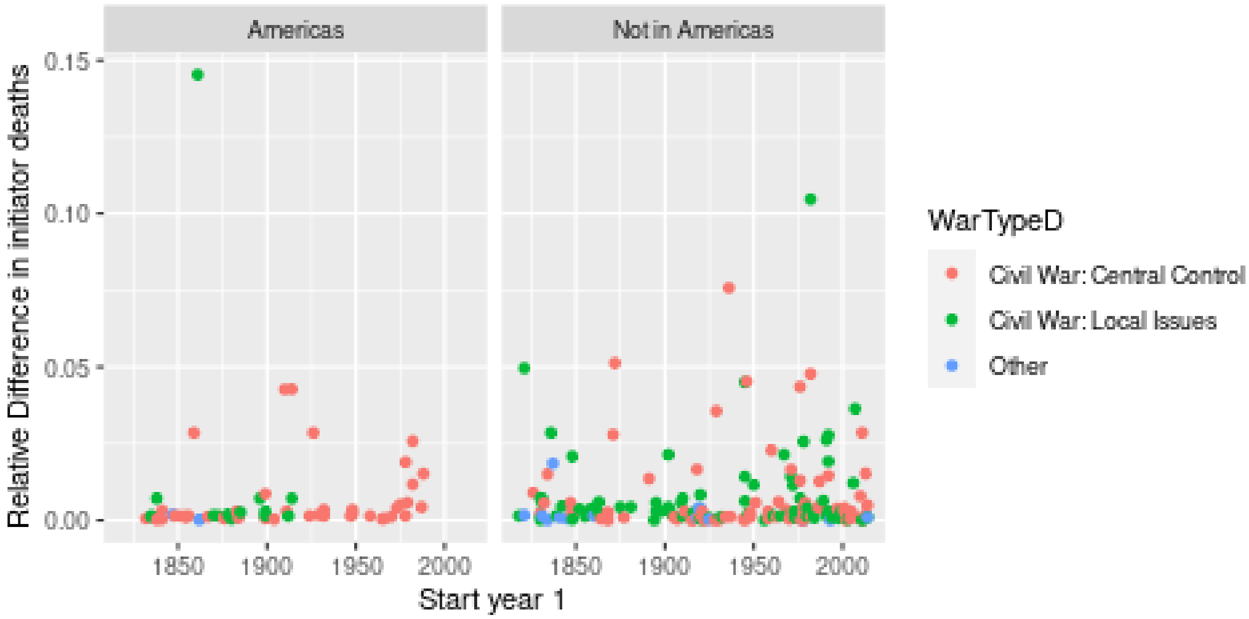
#Absolute difference in deaths Defined as the initiator deaths-recipient deaths, once several influential points (not outliers) were removed one can see a large proportion of initiator deaths more in the negatives from the 1900s onwards in the not in Americas group. Maybe this relates to some sort of 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(InitiatorDeaths)) 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 0.008
## 2
## 3
## 4
## 5
## 6
##                            Variables Estimate Std err p-val
## 1                        (Intercept) -5516.36 3845.83 0.154
## 2          ‘AmericasNot in Americas‘  7053.53 3342.07 0.037
## 3                       StartYR_Norm    18.92   25.39 0.458
## 4                         WDuratDays    -0.17    1.37 0.904
## 5 ‘WarTypeDCivil War: Local Issues‘ -5217.67    2974 0.082
## 6                      WarTypeDOther -2449.43 5883.64 0.678
```

```
##                                  Description    R2
## 1 Relative difference in deaths non-imputed 0.006
## 2
## 3
## 4
## 5
## 6
##                            Variables Estimate Std err  p-val
## 1                        (Intercept)    0.007   0.004   0.14
## 2          ‘AmericasNot in Americas‘   -0.003   0.004   0.41
## 3                       StartYR_Norm        0       0   0.29
## 4                         WDuratDays        0       0 <0.001
## 5 ‘WarTypeDCivil War: Local Issues‘    0.003   0.003   0.36
## 6                     WarTypeDOther   -0.003   0.007   0.64
```

#Models deaths here The common theme across models was prediction was very low between training and test sets. An association, however, was seen in the war type (‘local issues’). 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                         Variables
## 1 Absolute difference in deaths imputed 0.002                        (Intercept)
## 2                                                      ‘AmericasNot in Americas‘
## 3                                                                   StartYR_Norm
## 4                                                                     WDuratDays
## 5                                              ‘WarTypeDCivil War: Local Issues‘
## 6                                                                  WarTypeDOther
##   Estimate Std err p-val
## 1 -1189.23 2817.62 0.673
## 2  3308.69 2918.29 0.258
## 3     0.97   20.39 0.962
## 4     -0.2    0.88 0.825
## 5  -5899.8 2546.22 0.021
## 6   777.01 3817.33 0.839
```

```
##                             Description    R2                         Variables
## 1 Relative difference in deaths imputed 0.006                        (Intercept)
## 2                                                      ‘AmericasNot in Americas‘
## 3                                                                   StartYR_Norm
## 4                                                                     WDuratDays
## 5                                              ‘WarTypeDCivil War: Local Issues‘
## 6                                                                  WarTypeDOther
##   Estimate Std err p-val
## 1    0.003   0.004  0.46
## 2    0.002   0.004  0.64
## 3        0       0  0.38
## 4        0       0  0.12
## 5    0.005   0.003  0.15
## 6        0   0.005  1.00
```

#Internationalized Non-imputed

Network

```
##                                      Description   R2 Estimate Std err p-val
## 1 Intl Absolute difference in deaths non-imputed <NA> -1189.23 2817.62  0.67
## 2                                                      3308.69 2918.29  0.26
## 3                                                         0.97   20.39  0.96
## 4                                                         -0.2    0.88  0.83
## 5                                                      -5899.8 2546.22  0.02
## 6                                                       777.01 3817.33  0.84
```

```
##                                      Description   R2 Estimate Std err p-val
## 1 Intl Relative difference in deaths non-imputed <NA>   -0.006   0.012  0.61
## 2                                                            0    0.01  0.97
## 3                                                            0       0  0.87
## 4                                                            0       0     0
## 5                                                       -0.001   0.009   0.9
## 6                                                       -0.001    0.02  0.98
## 7                                                          0       0   0.9
## 8                                                          0       0  0.08
```

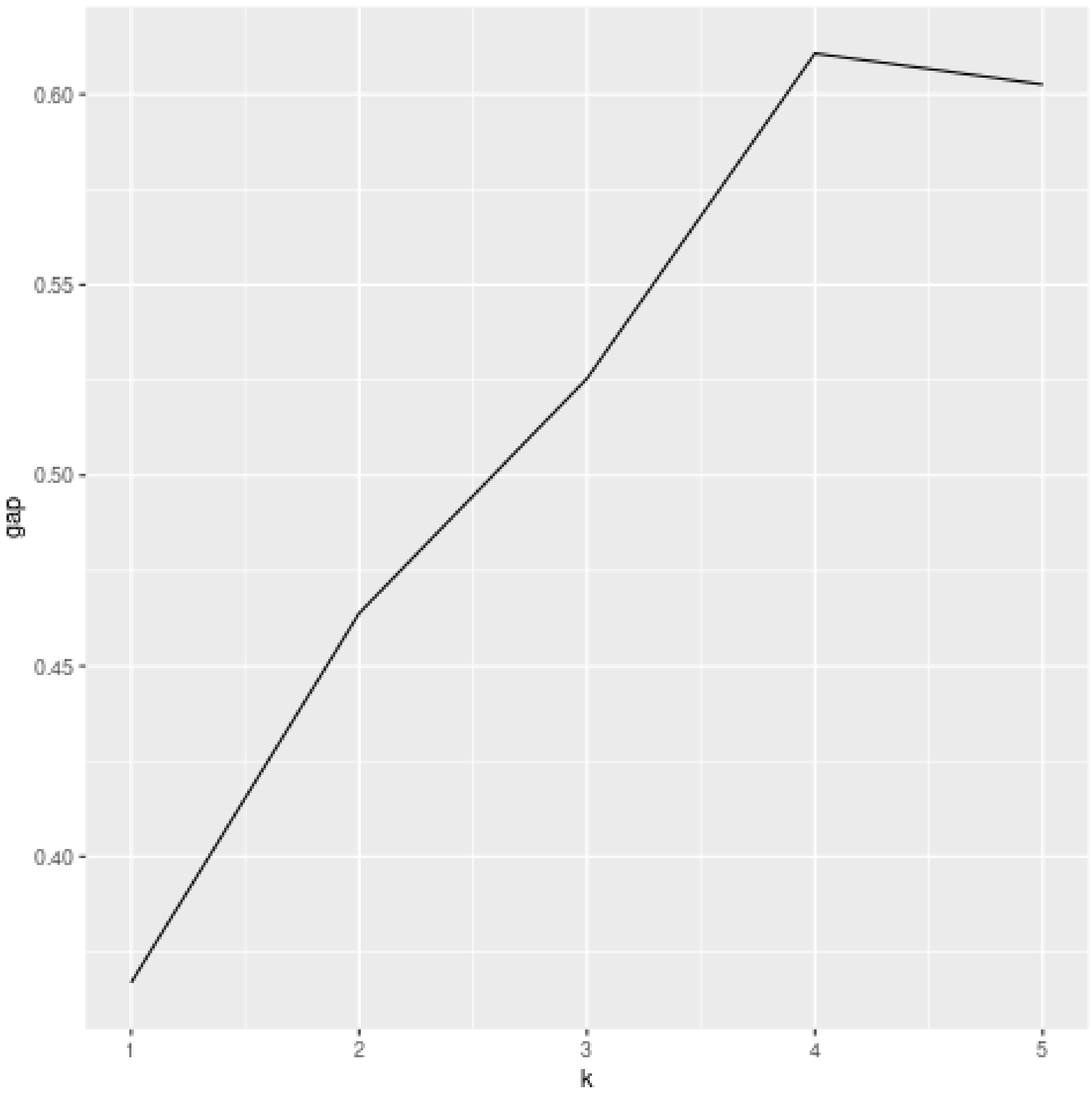
#Internationalized Imputed Network

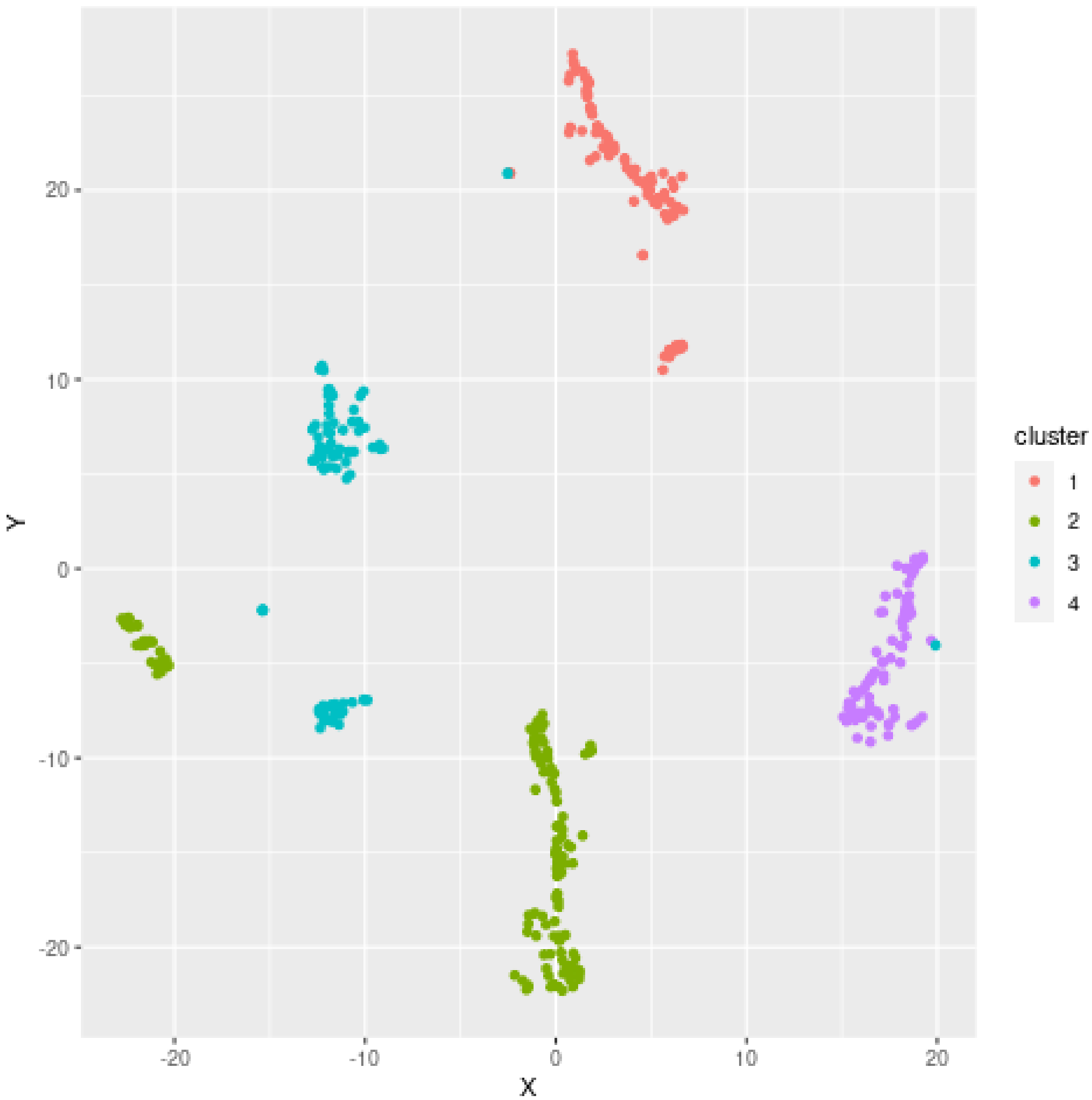
```
##                                      Description   R2 Estimate Std err p-val
## 1 Intl Absolute difference in deaths non-imputed <NA> -1189.23 2817.62  0.67
## 2                                                      3308.69 2918.29  0.26
## 3                                                         0.97   20.39  0.96
## 4                                                         -0.2    0.88  0.83
## 5                                                      -5899.8 2546.22  0.02
## 6                                                       777.01 3817.33  0.84
```

```
##                                      Description   R2 Estimate Std err p-val
## 1 Intl Relative difference in deaths non-imputed <NA>   -0.006   0.012  0.61
## 2                                                            0    0.01  0.97
## 3                                                            0       0  0.87
## 4                                                            0       0     0
## 5                                                       -0.001   0.009   0.9
## 6                                                       -0.001    0.02  0.98
## 7                                                            0       0   0.9
## 8                                                            0       0  0.08
```

#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



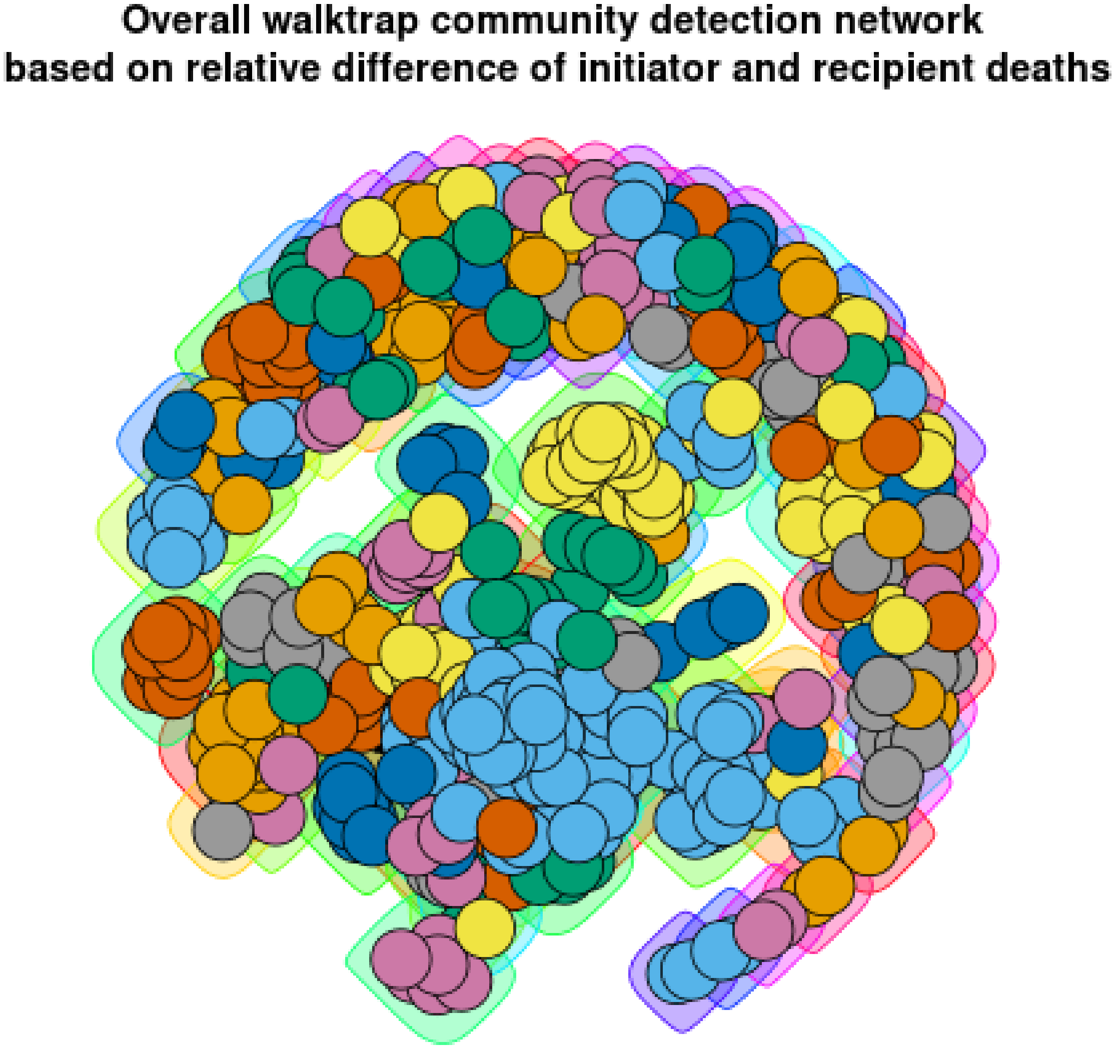
#Overall clustering by variables

As described above the four clusters showed a distince 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.

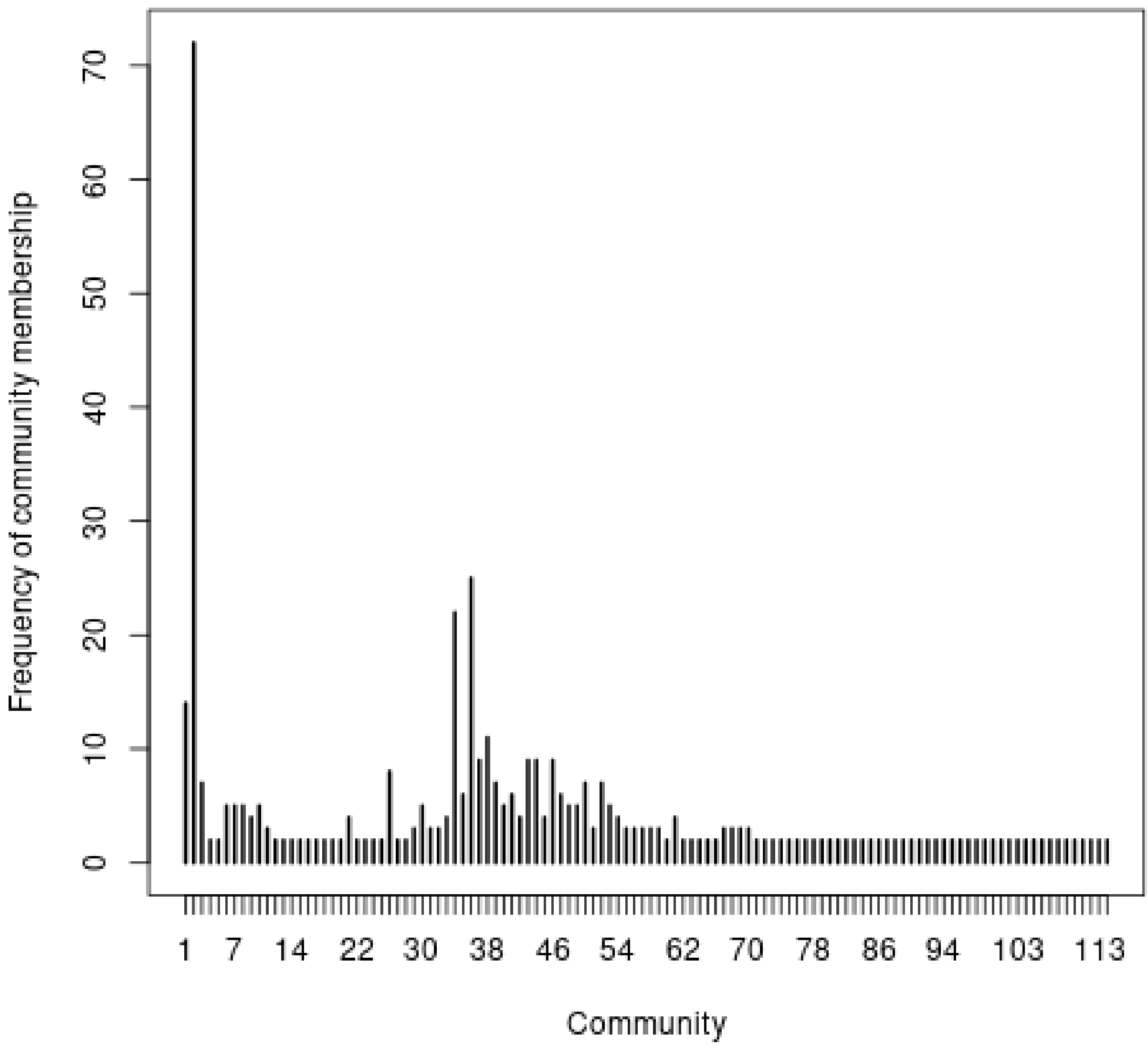
```
##
## --------Summary descriptives table by 'cluster'---------
##
## ----------------------------------------------------------------------------------------------------
##                                               1                2               3               4
##                                              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
## -----------------------------------------------------------------------------------------------------
```

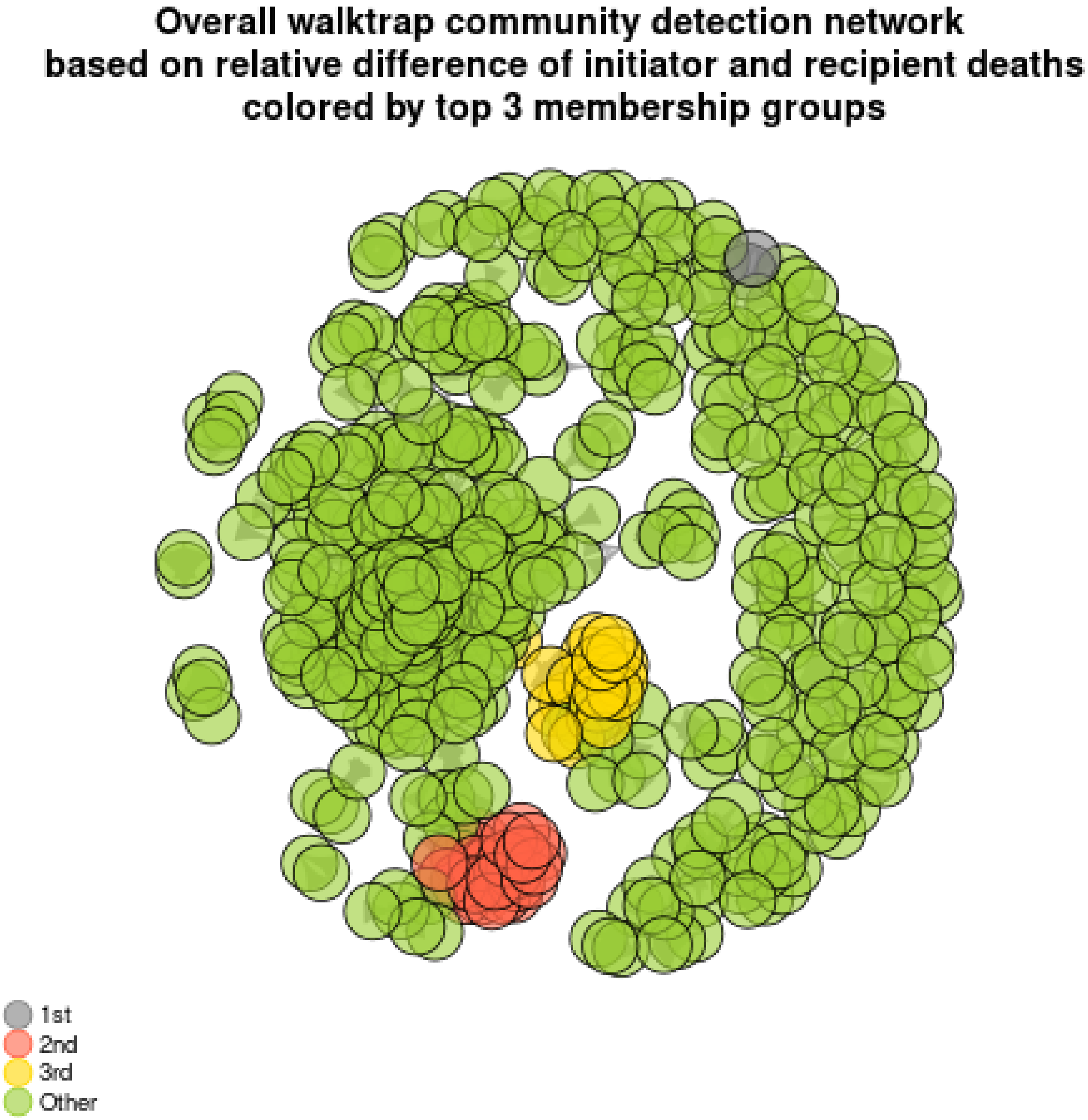
#Aim 4 network composition While not much was shown, several network plots were constructed. A walk trap community search algorithm was implemented and is generally a solid approach when not much is known about the given network structure. Weights between conflicts were normalized differences in conflicts 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 will be evaluated as a next-to-do.



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



Overall walktrap community detection network based on relative difference of initiator and recipient deaths colored by top 3 membership groups

#Conclusions

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 was the most interesting result as of now. 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 is still in process and other approaches could be considered that also try to look at the relationships in a flow of relationships.