Explanatory Memo for Generating Data and Replicating Analyses used in:
Who Wrote the Rules for the Trans-Pacific Partnership?

This memo details how to generate the data and replicate the analyses used in *Who Wrote the Rules for the Trans-Pacific Partnership?* Although we provide a user-friendly Excel file that contains all of the data returned from our text-comparison analyses (***TPP_data_RP.xls***), we first provide instructions for how to create these primary text-comparison variables (if desired). We then discuss the analyses that were done using these primary variables, which are reported in the manuscript in Figures 1-4 and Tables 1-2. Finally, we also include the relevant code used to generate the figures in the manuscript. The data for doing so are contained in the files "***TPP_data_RP.csv***" and "***heatmap_data_rp.csv***."

1) File that contains the Text-Comparison Variables

The included file (*TPP_data_RP.xls*) is the master file that contains all of the relevant results from our text-comparison analyses, described in pp. 7-15 of the paper.
You can just skip to section 3) below unless you are interested in generating these data on your own, the process for which is described in section 2a) below.

2) Generating the Text-Comparison Variables (optional)

a) Variables used in Figures 2-4 and Tables 1-2

All of the relevant data is contained in the provided file (*TPP_data_RP.xls*) and the other version of that file (*TPP_data_RP.csv*). Below are the instructions for generating those data, if desired.

One can generate the raw text-comparison data used in this paper using Wcopyfind 4.1.4 (Bloomfield 2014), along with the full collection of PTA texts (and chapters of PTAs texts), which are available at:
https://www.dropbox.com/sh/sdjbtcxuelsfjqw/AABPkL4FBzYX012uDUzQKkpza?dl=0.
The text files are contained in a set of 16 folders, all of which are contained within a folder "*RP Text Files*." The sub-folder "*Full Agreements*" contains the complete text of the TPP and the 74 PTAs. The other 15 sub-folders are labeled by issue area (*e.g.,* "Antidumping", "Dispute Settlement") and each sub-folder contains one text file with the TPP text and other text files with the parallel chapters from PTAs. In order to generate our primary indicators of how much matching text there is between the TPP and PTAs, we use the program Wcopyfind 4.1.4, which runs as a single executable (.exe) file for windows and can be found here:

http://plagiarism.bloomfieldmedia.com/wordpress/software/wcopyfind/.

The program allows the user to specify a corpus of documents, adjust match parameters, and run comparisons between documents. For our comparisons we set the parameters to the recommended settings, which is the standard used in other political science studies. More precisely, once the GUI launches we set the match parameters to 6 words, with the fewest matches to report as 1 word. Further, we chose the following options: "ignore all punctuation," "ignore outer punctuation," "ignore numbers," "ignore letter case," and "skip non-words." In order to generate the comparisons all documents must be loaded as either .txt or .doc files. We provide two sets of .txt files. The full-length TPP and PTA texts are in the folder labeled "*Full Agreements*" whereas the chapter texts are separated in folders by issue area.

To set up and execute a comparison, place the .txt files of the relevant PTAs (or PTA chapters) in the top window, which is called "Old Document Files." Then place the .txt file of the TPP (or TPP subsections) in the bottom window, which is called "New Document Files." The TPP files will always begin with 001. Click run.

After the analysis is complete the program returns a list, automatically as an .html window in Internet Explorer, which contains five columns. The first column, labeled "perfect match," is a calculation of the percentage match for pairs of documents in the set, where the percentage match is the total number of matching words shared between the two documents divided by the total words in each document. The column contains three things: 1) the number of jointly shared words between the two documents being compared; 2) the percentage of words copied in the 1st (left) document, and; 3) the percentage of shared words copied in the 2nd (right) document. The second column contains overall match numbers, which are the numbers Wcopyfind generates if the user specifies error parameters for the text comparisons. The third column allows the user to compare the text of the two agreements side-by-side in a new window. Finally, the fourth and fifth columns indicate the two documents that were compared, labeling them left and right. We primarily use numbers from the first column since this indicates perfectly matching text shared between the two documents specified as left and right.

i) From these analyses we create a primary variable that measures the extent of textual overlap between the PTA and the TPP, based on the entire contents of both. This variable, *pm_percent_pta*, is the basis for all calculations presented and discussed in pp. 7-11 (Figures 2-3 and Table 1). It is presented in column R in the attached Excel file (*TPP_data_RP.xls*).

ii) For the chapter-specific analyses (in Table 2 and Figure 4), we compare the texts from each of the 15 chapters using the same procedures described above. These analyses use the text files contained in each of the 15 folders provided (see Figure A), and generate parallel variables for each chapter for the amount of text overlap between the PTA and TPP.

Figure A. Summary of Wcopyfind Variables

Trade Issue Area	Variables	From Folder
Perfect Match	*pm_percent_pta*	*Full Agreements*
Overall Match	*om_percent_pta*	*Full Agreements*
Antidumping	*ad_pm_words_pta*	*Antidumping*
Dispute Settlement	*ds_pm_percent_pta*	*Dispute Settlment*
E-Commerce	*ecom_pm_percent_pta*	*E-Commerce*
Environment	*env_pm_percent_pta*	*Environment*
Financial Services	*fin_pm_percent_pta*	*Financial*
Investment	*invs_pm_percent_pta*	*Investment*
Intellectual Property	*ipr_pm_percent_pta*	*Intellectual Property*
Movement of Persons	*move_words_pta*	*Movement*
Labor	*lab_pm_percent_pta*	*Labor*
Procurement	*proc_pm_percent_pta*	*Procurement*
Safeguards	*safe_pm_percent_pta*	*Safeguards*
Services	*serv_pm_percent_pta*	*Services*
Sanitary and Phyto-Sanitary	*sps_pm_percent_pta*	*Sanitary and Phyto-Sanitary*
Technical Barriers	*tbt_pm_percent_pta*	*TBTs*
Telecommunications	*tele_pm_percent_pta*	*Telecom*

b) Variables used in heat map in Figure 1

Using the methodology described above in section 2a), we also compare the 74 pre-existing PTAs of TPP members (1995-2015) to each other. The results of these comparisons are contained in the included file "*heatmap_data_rp.csv.*"

The data for constructing the heatmap is created in a similar manner to the TPP-PTA comparisons described in 2a) above. The primary difference, however, is that for these comparisons every pairwise PTA comparison must be obtained. In other words, all 74 PTAs are compared to every other PTA. In order to do this, the PTAs are loaded in the "new document files" window in WcopyFind, which then generates a list of the 2738 relevant PTA comparisons. These then can be transformed in R to obtain a 74x74 symmetric matrix required for the heatmap. For these PTA-PTA matches we use the same parameters as in the main analyses described above (the percentage overlap numbers from the more conservative perfect match comparisons). The results overlap percentage numbers are depicted in the variable "perfectmatch_percent_doc1" in the "*heatmap_data_rp.csv*" file. The doc1 and doc 2 files contain the relevant labels.

The full R script for generating the map from the raw data is contained in the r-markdown file included in section 4 of this memo.

3) Primary Calculations (those in Figures 2-4 and Tables 1-2)

The data presented in Figures 2-3 and Table 1 are drawn from the above-described analyses of entire agreements, while Figure 4 and Table 2 are taken from the above-described analyses of the chapters. Refer back to Part 2, and the files described there, for details. Everything described in this section is highlighted in the enclosed Excel file (*TPP_data_RP.xls*).

a) Figure 2 depicts the degree to which each PTA matches the TPP based on *pm_percent_pta*. See column R in the included Excel file (*TPP_data_RP.xls*).

b) Figure 3 presents country-averages based on *pm_percent_pta*. The averages are computed in in the included Excel file and are highlighted in orange in column R.

c) Table 1 contains robustness checks for the country-level averages contained in Figure 3. All of the information in columns 3-7 of Table 1 also are included in the include Excel file (*TPP_data_RP.xls*), in columns S-Z. The country averages are highlighted in yellow.

d) Table 2 summarizes the countries and PTAs with the greatest match with the TPP across the 15 different chapters. The 15 relevant columns in the Excel file are those that take the form of *chaptername_pm_percent_pta.* For each of the 15 columns, the values highlighted in red represent the PTA with the greatest match with the TPP for that chapter and the values in yellow represent the largest country average for that chapter.

e) Figure 4 plots for the investment chapters: i) the average match for each country with the TPP investment chapter, and ii) the 10 investment chapters in PTAs that overlap with the most with the investment chapter in the TPP. These numbers are drawn from column CA in the Excel file.

4) R Code for Generating Figures

This next section is an r-markdown file that includes the syntax used to create all of the figures presented in the manuscript (Figures 1-4). We also have included the R script (*TPP Paper_RP.r*). Comments are preceded by a "##" whereas usable code is not.

Creating Figure 1

```
####Heatmap code (for figure 1 in manuscript)
##Bring in data
Data_frame <- read.csv("~heatmap_data_rp.csv", stringsAsFactors = FALSE
)

##Specify the relevant variables
```

```
keep <- c("doc1", "doc2", "perfectmatch_percent_doc1")
df_new <- Data_frame[keep]

##change order of list here before convering to matrix
final_list<- c("999", "543","84", "96", "187", "188", "218", "241", "55
1", "628", "637", "643", "645", "658", ##US
               "520", "518", "517", "519", "146", "207", "495", "523",
"826", "829", "842", ##Japan
               "164", "165", "168", "163", "162","161", "166", "798", "
797", "852",  ##Canada
               "82", "83", "75", "795", "911", "950", ##Australia
               "390", "509", "330", ##Mexico
               "599", "598", "800", "951", "953", ##Malaysia
               "631", "396", "534", "493", "550", "641", "228", "475",
"644", "814", "874", ##singapore
               "202", "205", "208", "199", "206", "802", "858", ##chile
               "771", "227", "819", "830", "810", ##peru
               "632", "222", "825") ##new zealand


##Put the list into matrix
nm_final<-matrix(NA, nrow=length(final_list), ncol=length(final_list),
dimnames=list(final_list, final_list))

##now fill the matrix

for(i in 1:length(df_new[,1])){
  nm_final[grep(paste(df_new[i,1]),rownames(nm_final)), grep(paste(df_n
ew[i,2]),colnames(nm_final))]<-df_new[i,3]
  nm_final[grep(paste(df_new[i,2]),rownames(nm_final)), grep(paste(df_n
ew[i,1]),colnames(nm_final))]<-df_new[i,3]
}

diag(nm_final) = 1
View(nm_final)
isSymmetric(nm_final)

## [1] TRUE

##apply new rownames
r1<-c("TPP", "US Jordan","US Australia", "US Bahrain",  "CAFTA", "CAFTA
-DR", "US Chile", "US Colombia", "US Korea", "US Morocco", "US Oman", #
US
      "US Panama", "US Peru", "US Singapore", #US
      "Japan Singapore", "Japan Mexico", "Japan Malaysia",  "Japan Phil
ippines",  "Japan Brunei",  "Japan Chile",  "Japan Indonesia", #Japan
      "Japan Vietnam", "Japan India", "Japan Peru", "Japan Australia",
#Japan
      "Canada EFTA","Canada Israel", "Canada Peru", "Canada Costa Rica"
, "Canada Colombia", "Canada Chile", "Canada Jordan", "Canada Panama",
```

```
"Canada Honduras", "Canada Korea", #Canada
      "Australia Singapore", "Australia Thailand", "Australia Chile", "
Australia Malaysia", "Australia Korea", "ASEAN-Aus-NZ", #Australia
      "Mexico EFTA", "Mexico Israel", "Mexico EC", #Mexico
      "Malaysia Pakistan", "Malaysia New Zealand", "Malaysia Chile", "M
alaysia India", "Malaysia Turkey", #Malaysia
      "Singapore NZ", "Singapore EFTA",  "Singapore Jordan", "Singapore
 India", "Singapore Korea", "Singapore Panama", "Singapore China", "Sin
gapore GCC", "Singapore Peru", "Singapore Costa Rica", "Singapore EC",
#Singapore
      "Chile EC", "Chile EFTA", "Chile Korea", "Chile China", "Chile In
dia", "Chile Turkey", "Chile Hong Kong", #Chile
      "Peru Thailand", "Peru China", "Peru EFTA", "Peru Korea", "Peru C
ol-EC", #Peru
      "NZ Thailand", "NZ China", "NZ Hong Kong") #New Zealand

rownames(nm_final)<-r1
colnames(nm_final)<-r1

#melt for ggplot
melted <- melt(nm_final)
head(melted)

##Code for final heatmap figure 1 in manuscript
ggplot(data = melted, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_distiller(palette = "YlOrRd", direction = 1) +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90, vjust = 1,
                                   size = 4, hjust = 1, family="Calibri
", face="bold"),
        axis.text.y = element_text(size = 4, family="Calibri", face="bo
ld"),
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        panel.grid.major = element_blank(),
        panel.border = element_blank(),
        panel.background = element_blank(),
        axis.ticks = element_blank(),
        legend.justification = c(.9, 0),
        legend.position = c(1.15, 0.35),
        legend.direction = "vertical",
        legend.title=element_text(size=7, family="Calibri", face="bold"
),
        legend.text=element_text(size=6, family="Calibri", face="bold")
) +
  guides(fill = guide_colorbar(barwidth = 1, barheight = 4))+
  ggtitle("") +
  theme(plot.title = element_text(face="bold", size=8, hjust=0, vjust=0
```

```
)) +
  coord_fixed()
```

Background Steps for Figures 2-4

```
##R script for Research and Politics Manuscript
##Who Wrote the Rules for the Trans-Pacific Partnership

library(ggplot2)
library(multcomp)

library(gplots)

library(coefplot)
library(DTK)
library(plyr)
library(reshape2)
library(gridExtra)
library(grid)

##Bring in the data
df <- read.csv("~/TPP_data_RP.csv", stringsAsFactors = FALSE)

# sort data in order of percent overlap with TPP
df_sorted <- df[order(df$pm_percent_pta, decreasing=T), ]
##Sort Data by agreement for Figure 2
df_sorted$name <- reorder(df_sorted$name, df_sorted$pm_percent_pta)
df_sorted$country <-reorder(df_sorted$country, df_sorted$pm_percent_pta
)
```

Creating Figure 2

```
##Figure 2 -- percent match by agreement
ggplot(df_sorted, aes(x = pm_percent_pta, y = factor(name))) +
  geom_point(aes(color = factor(usa_dum))) +
  scale_color_manual(values=c("black", "blue")) +
  theme_bw() +
  theme(axis.title.x = element_text(size = 12, vjust = .25),
        axis.text.y = element_text(size = 4),
        axis.title.y = element_blank(),
        legend.text= element_text(size=6),
        legend.title=element_text(size=6),
        legend.position = "none") +
  xlab("Percent Overlap")  +
  ggtitle("Textual Overlap Between PTAs and TPP")

##Looking at country level patterns
avg<-aggregate( formula = pm_percent_pta~country,
                data = df,
```

```
                FUN = mean )


avg$country <-reorder(avg$country, avg$pm_percent_pta)

avg[order(avg$country,decreasing = TRUE),]
```

Test of Country Differences

```
##Assessing Statistical significance of country differences

##Step 1, Performing ANOVA for equal and unequal variances
oneway.test(df$pm_percent_pta ~ df$country)

##
##  One-way analysis of means (not assuming equal variances)
##
## data:  df$pm_percent_pta and df$country
## F = 16.067, num df = 11.000, denom df = 22.151, p-value =
## 4.347e-08

oneway.test(df$pm_percent_pta ~ df$country, var.equal = TRUE)

##
##  One-way analysis of means
##
## data:  df$pm_percent_pta and df$country
## F = 8.5515, num df = 11, denom df = 89, p-value = 4.143e-10

##Performing TukeyHSD test for group differences
m<-aov(df$pm_percent_pta ~ df$country)
summary(m)

##             Df Sum Sq Mean Sq F value   Pr(>F)
## df$country  11   3727   338.9   8.551 4.14e-10 ***
## Residuals   89   3527    39.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(m)

#US mean statistically different than all other group means; p<.05
```

Creating Figure 3

```
##Visualizing the data
##Put the Data in order and Label US
avg$country <-reorder(avg$country, avg$pm_percent_pta)
us <- subset(avg, country == "USA")
round<-round(avg$pm_percent_pta, digits=0)
```

```
round<-sort(round, decreasing=TRUE)

##Figure 3 in manuscript
ggplot(avg, aes(x = pm_percent_pta, y = factor(country))) +
  geom_point(color = "black") +
  geom_point(data=us, color="blue") +
  theme_bw() +
  theme(axis.title.x = element_text(size = 10, vjust = .25, family="Cal
ibri", face="bold"),
        axis.text.y = element_text(size = 10, family="Calibri", face =
"bold"),
        axis.title.y = element_blank(),
        legend.position = "none") +
  #geom_text(aes(label=round, vjust=-.35)) +
  xlab("Percent Replicated in TPP")  +
  ggtitle("")

##Assessing robustness of patterns -- Table 1 in manuscript
##Note: Some calculations done in excel

##This calculates match % with relaxed parameters
avg_relaxed<-aggregate( formula = om_percent_pta~country,
                            data = df,
                            FUN = mean )

avg_relaxed$country <-reorder(avg_relaxed$country, avg_relaxed$om_perce
nt_pta)
avg_relaxed[order(avg_relaxed$country,decreasing = TRUE),]
```

Creating Figure 4

```
##Making the investment Plots
###Investment - individual agreements
invs_sort <- df[order(df$invs_pm_percent_pta, decreasing=T), ]
invs_sort <- invs_sort[1:13,]
invs_sort$name <- reorder(invs_sort$name, invs_sort$invs_pm_percent_pta
)
us_invs <- subset(invs_sort, country == "USA")
invs <- ggplot(invs_sort, aes(x = invs_pm_percent_pta, y = factor(name)
)) +
  geom_point() +
  geom_point(data=us_invs, color="blue") +
  scale_x_continuous(limits = c(75, 90))+
  theme_bw() + theme(axis.title.x = element_text(size = 10, vjust = .25
,family="Calibri", face="bold"),
                     axis.text.y = element_text(size = 10, family="Cali
bri", face="bold"),
                     axis.title.y = element_blank(),
                     plot.title = element_text(size=12, family="Calibri
```

```
", face="bold"))+
  xlab("Percent Replicated in TPP") + ggtitle("10 Closest-Match Agreeme
nts")

##Investment -- country averages
avg<-aggregate( formula = invs_pm_percent_pta~country,
                data = df,
                FUN = mean )

avg_sorted <- avg[order(avg$invs_pm_percent_pta, decreasing=T), ]
avg_sorted$country <- reorder(avg_sorted$country, avg_sorted$invs_pm_pe
rcent_pta)
us_invs_avg <- subset(avg_sorted, country == "USA")
invs_avgs <- ggplot(avg_sorted, aes(x = invs_pm_percent_pta, y = factor
(country))) +
  geom_point() +
  scale_x_continuous(limits=c(20, 80))+
  geom_point(data=us_invs_avg, color="blue") +
  theme_bw() + theme(axis.title.x = element_text(size = 10, vjust = .25
, family="Calibri", face="bold"),
                     axis.text.y = element_text(size = 10, family="Cali
bri", face="bold"),
                     axis.title.y = element_blank(),
                     plot.title = element_text(size=12, family="Calibri
", face="bold"))+
  xlab("Percent Replicated in TPP") + ggtitle("Country Averages")

##Putting the two plots together; Figure 4 in Manuscript
grid.arrange(invs_avgs, invs, ncol=2) ##use 6.5 x 3 for dimensions
```