DNM vs Alignment Comparison

Jake Prasad

Tuesday, July 28, 2015

```
## Warning: package 'bit64' was built under R version 3.1.3

## Loading required package: bit
## Attaching package bit
## package:bit (c) 2008-2012 Jens Oehlschlaegel (GPL-2)
## creators: bit bitwhich
## coercion: as.logical as.integer as.bit as.bitwhich which
## operator: ! & | xor != ==
## querying: print length any all min max range sum summary
## bit access: length<- [ [<- [[ [[<-
## for more help type ?bit
##
## Attaching package: 'bit'
##
## The following object is masked from 'package:data.table':
##
##     setattr
##
## The following object is masked from 'package:base':
##
##     xor
##
## Attaching package bit64
## package:bit64 (c) 2011-2012 Jens Oehlschlaegel (GPL-2 with commercial restrictions)
## creators: integer64 seq :
## coercion: as.integer64 as.vector as.logical as.integer as.double as.character as.bin
## logical operator: ! & | xor != == < <= >= >
## arithmetic operator: + - * / %/% %% ^
## math: sign abs sqrt log log2 log10
## math: floor ceiling trunc round
## querying: is.integer64 is.vector [is.atomic} [length] is.na format print
## aggregation: any all min max range sum prod
## cumulation: diff cummin cummax cumsum cumprod
## access: length<- [ [<- [[ [[<-
## combine: c rep cbind rbind as.data.frame
## for more help type ?bit64
##
## Attaching package: 'bit64'
##
## The following object is masked from 'package:bit':
##
##     still.identical
##
## The following objects are masked from 'package:base':
##
##     :, %in%, is.double, match, order, rank

## Warning: package 'dplyr' was built under R version 3.1.3
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:data.table':
##
##     between, last
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
smoothalign <- function(df,sm,align="logodds") {
  if (align=="logodds") {
    return(log(df$ba+sm)-log(df$ba+df$nba+2*sm)-log(df$bna+sm)+log(df$bna+df$nbna+2*sm))
  } else if (align=="subodds") {
    return((df$ba+sm)/(df$ba+df$nba+2*sm)-(df$bna+sm)/(df$bna+df$nbna+2*sm))
  } else if (align=="logdnm") {
    return(log(df$ba+sm)-log(df$ba+df$nba+2*sm)-log(df$bna+df$ba+sm)+log(df$nba+df$ba+df$bna+df$nbna+2*s
  } else if (align=="subdnm") {
    return((df$ba+sm)/(df$ba+df$nba+2*sm)-(df$bna+df$ba+sm)/(df$nba+df$ba+df$bna+df$nbna+2*sm))
  } else {
    stop("Invalid alignment type.")
  }
}

d$lo1 <- smoothalign(d,1,"logodds")
d$sd0 <- smoothalign(d,0,"subdnm")

stopifnot(max(abs(d$lo1-d$pyalign))<.00001)
stopifnot(max(abs(d$sd0-d$dnm))<.00001)
```

```
## Warning in max(abs(d$sd0 - d$dnm)): no non-missing arguments to max;
## returning -Inf
```

Alignment (Verified/NonVerified):

Our standard measure for alignment uses Verified/Not verified as a power proxy. Using Verified/Not verified, we're able to clearly see that an influential speaker/uninfluential replier pair results in a stronger alignment than an uninfluential speaker/influential replier pair.

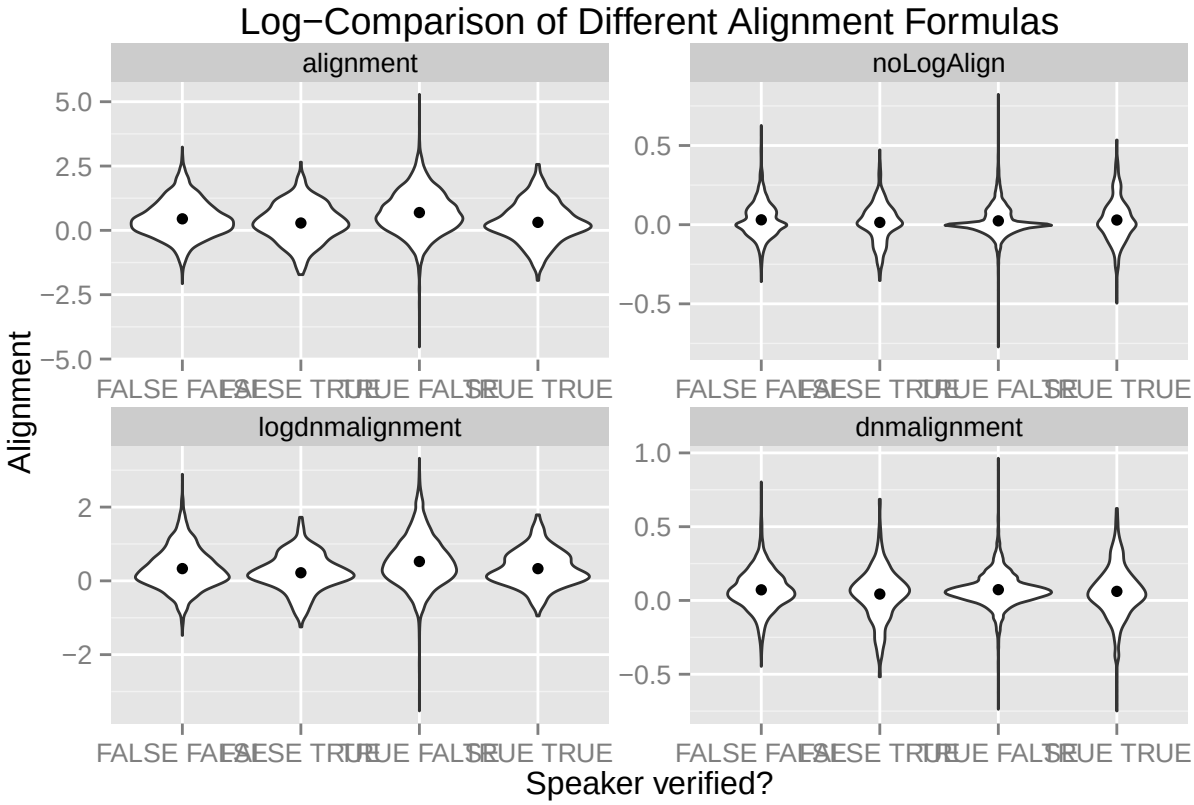
```
subsetted <- subset(df,logdnmalignment!="FALSE"&(ba+nba)>5&(bna+nbna)>5)
subsetted = transform(subsetted,logdnmalignment=as.numeric(logdnmalignment))
```

```
## Warning: NAs introduced by coercion
```

```
d2 <- subsetted %>%
  group_by(verifiedSpeaker,speakerId,replierId) %>%
  summarize(convs=n(), alignment=alignment, vreply=verifiedReplier, dnmalignment=dnmalignment, noLogAli
```

```
  gather(alignmentType,alignmentValue,c(alignment, dnmalignment, logdnmalignment, noLogAlign))

levels(d2$alignmentType)=c("alignment","noLogAlign","logdnmalignment","dnmalignment")

ggplot(d2,aes(x=paste(verifiedSpeaker,vreply),y=alignmentValue)) + geom_violin() + labs(title="Log-Compa
```

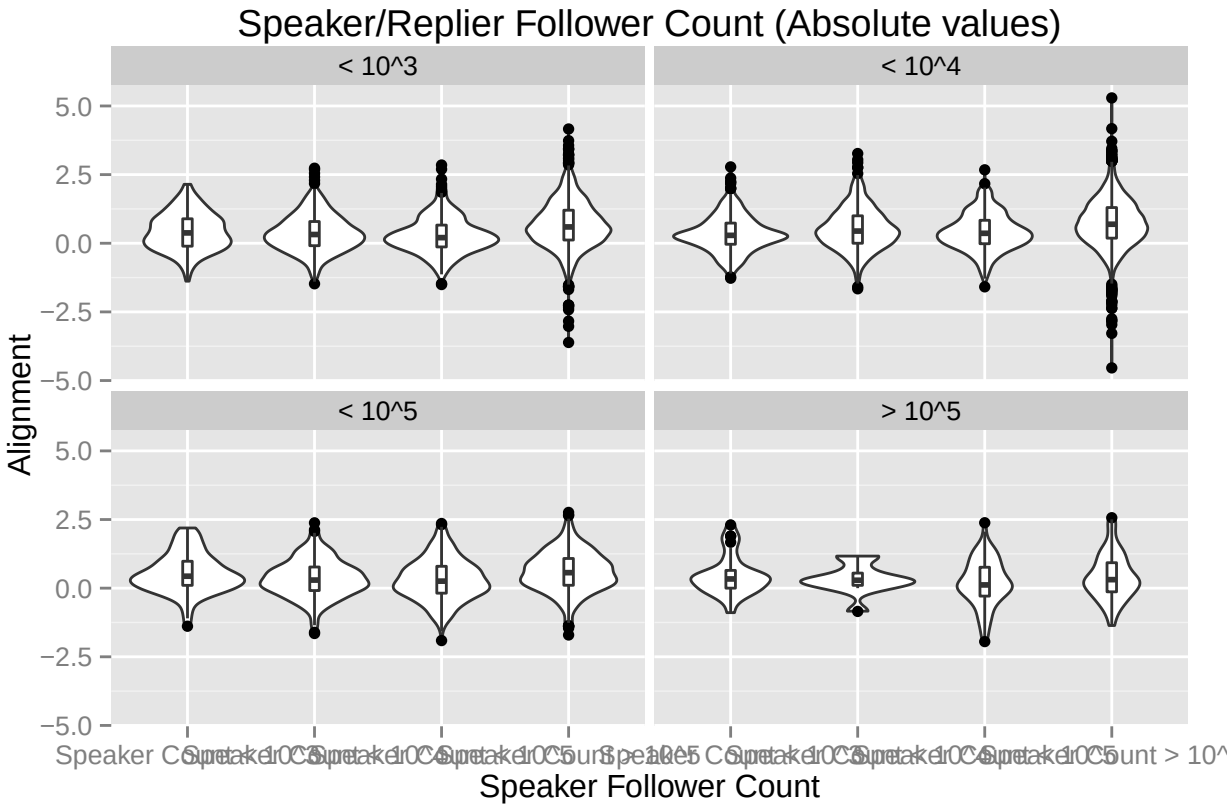
```
## Warning: Removed 4875 rows containing non-finite values (stat_ydensity).

## Warning: Removed 4875 rows containing missing values (stat_summary).
```



Follower Bins:

The Echoes Of Power paper used follower count as a power proxy. If follower count isn't a good power proxy it could explain why alignment wasn't found. We bin follower counts for speaker and repliers and plot the results.

```
d2 <- subset(df,logdnmalignment!="FALSE"&(ba+nba)>5&(bna+nbna)>5)
d2$speakerBins <- cut(d2$speakerFollowers, breaks=c(0,1000, 10000, 100000, 100000000), labels=c("Speake
d2$replierBins <- cut(d2$replierFollowers, breaks=c(0,1000, 10000, 100000, 100000000), labels=c("< 10^3
ggplot(d2,aes(x=paste(speakerBins),y=alignment)) + geom_violin() + labs(font=10, title="Speaker/Replier
```

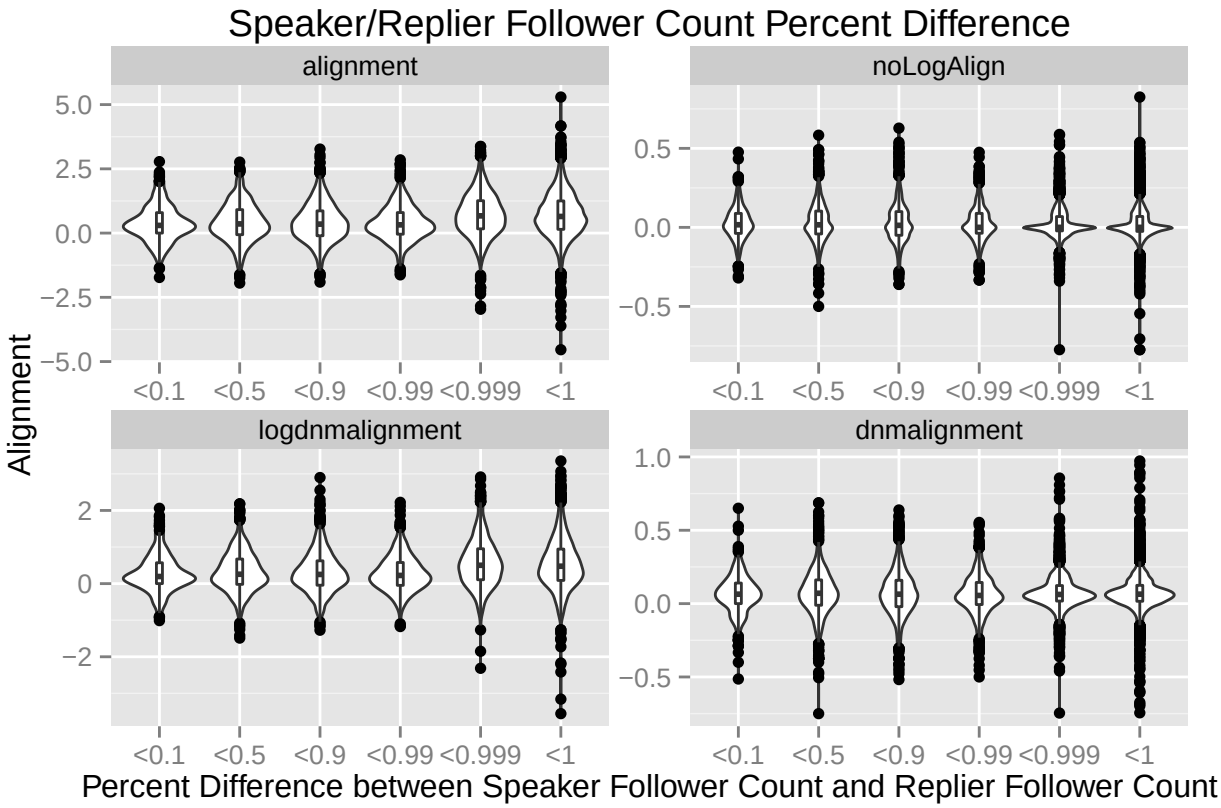


This is an interesting plot. It appears that follower count does have an effect on alignment. Let's also try plotting the percent difference between speaker follower count and replier follower count.

```
d2 <- subset(df,logdnmalignment!="FALSE"&(ba+nba)>5&(bna+nbna)>5)
subsetted = transform(subsetted,logdnmalignment=as.numeric(logdnmalignment))
d2 <- subsetted %>%
  group_by(verifiedSpeaker,speakerId,replierId) %>%
  summarize(convs=n(), alignment=alignment, vreply=verifiedReplier, dnmalignment=dnmalignment, noLogAlig
  gather(alignmentType,alignmentValue,c(alignment, dnmalignment, logdnmalignment, noLogAlign))

d2$followerBins <- cut(d2$percentDiff, breaks=c(0,0.1,0.5,0.9, 0.99, 0.999, 1), labels=c("<0.1","<0.5",

levels(d2$alignmentType)=c("alignment","noLogAlign","logdnmalignment","dnmalignment")

ggplot(d2,aes(x=followerBins,y=alignmentValue)) + geom_violin() + labs(title="Speaker/Replier Follower
```

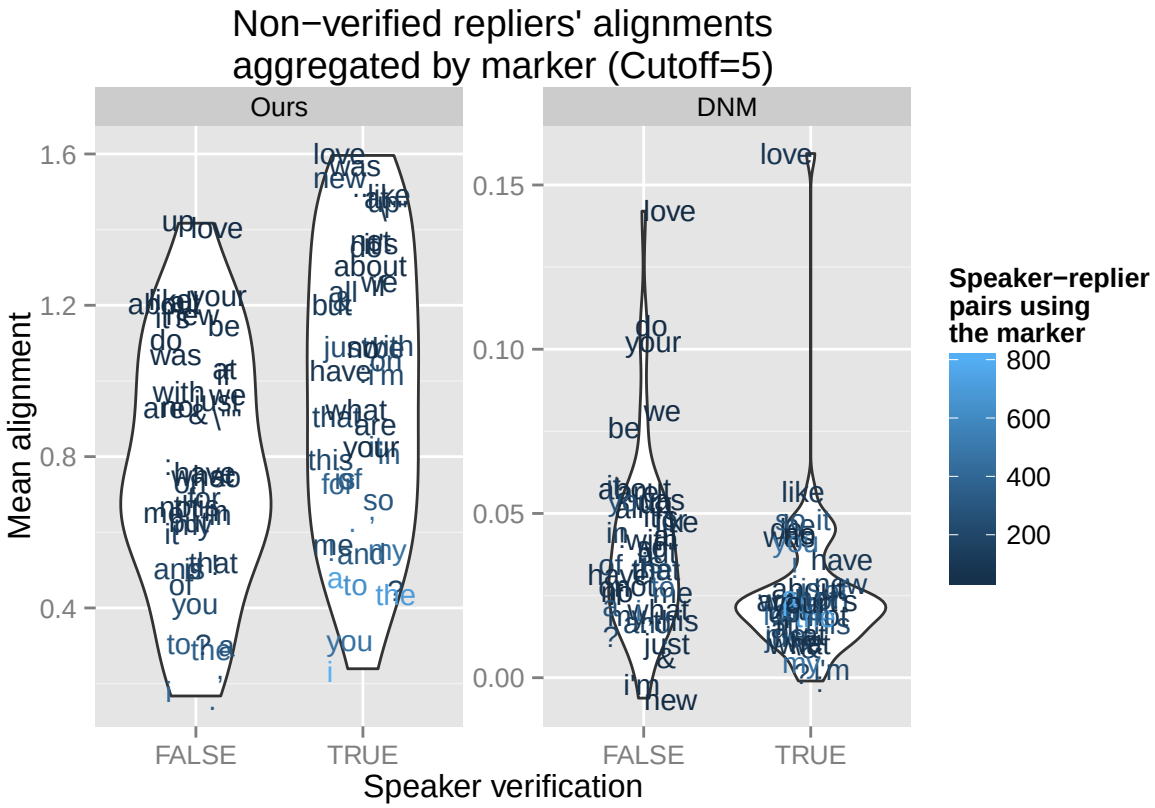
```
## Warning: Removed 4875 rows containing non-finite values (stat_ydensity).
```

```
## Warning: Removed 4875 rows containing non-finite values (stat_boxplot).
```



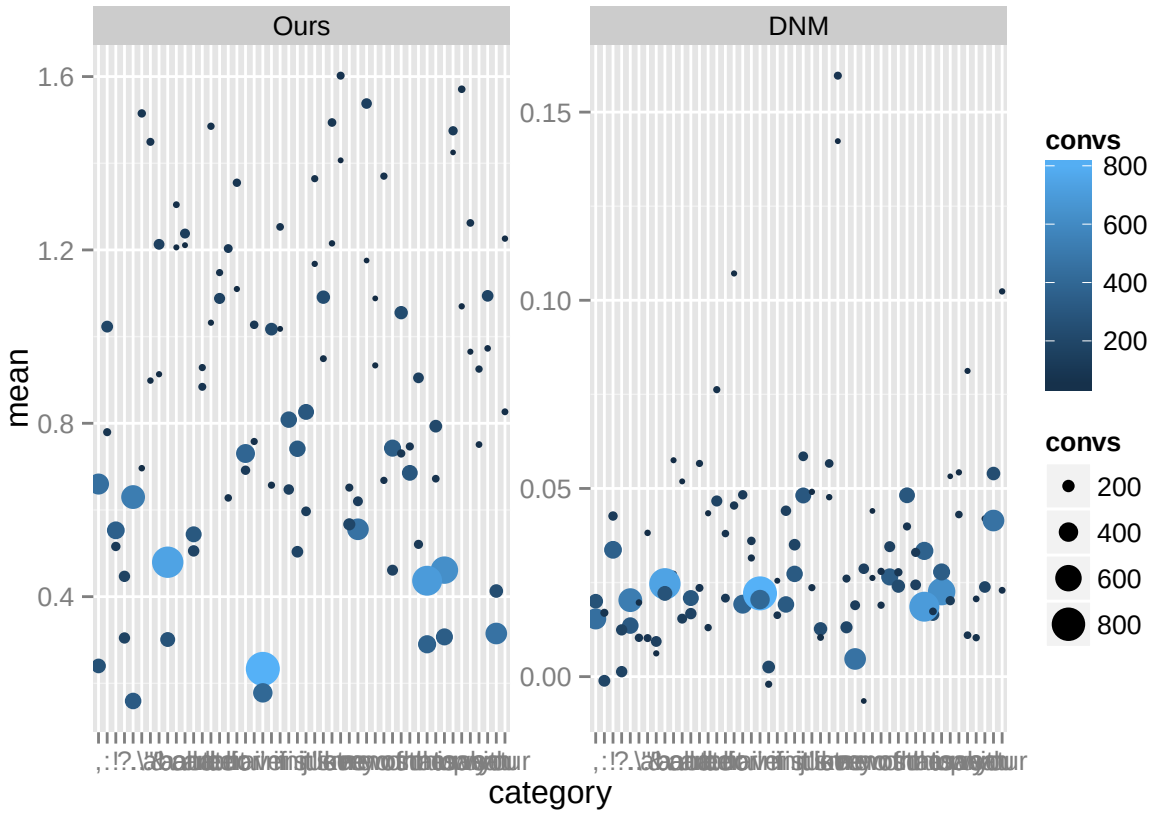
This is also interesting. It appears like alignment could follow a parabolic curve insteaed of a linear one. Let's also look at alignment by marker

```
d2 <- d %>%
  filter(vreply==F,(ba+nba)>=5,(bna+nbna)>=5) %>%
  group_by(vspeak,category) %>%
  summarize(convs=n(),Ours=mean(lo1),DNM=mean(sd0)) %>%
  gather(alignment,mean,Ours,DNM)

ggplot(d2,aes(x=vspeak,y=mean,color=convs,label=category)) + geom_violin() + geom_text(position=positio
```



Hmm, from the given plot, it seems the DNM formula is prone to outliers among individual markers. Why might this be the case? First we pot alignment values against number of conversations to make sure our intuition is correct

```
ggplot(d2,aes(x=category,y=mean,size=convs,color=convs,label=category)) + geom_point() + facet_wrap(~ali
```



Ok, this seems to indicate that alignment correlates to number of conversations. How might number of conversations affect alignment? If speakers don't use the marker in many conversations, the marker probably isn't very commonly spoken. Therefore a given person will have a low base probability of using the marker.

This seems to suggest that the DNM formula may be more strongly affected by a person's base rate, than their alignment.

We see a similar effect in the proposed alignment formula. However base rate doesn't appear to affect alignment as strongly, suggesting the proposed formula may be more accurate in calulating alignment.