Updates April 11

Owen G. Ward

4/11/2022

Poisson Initialization

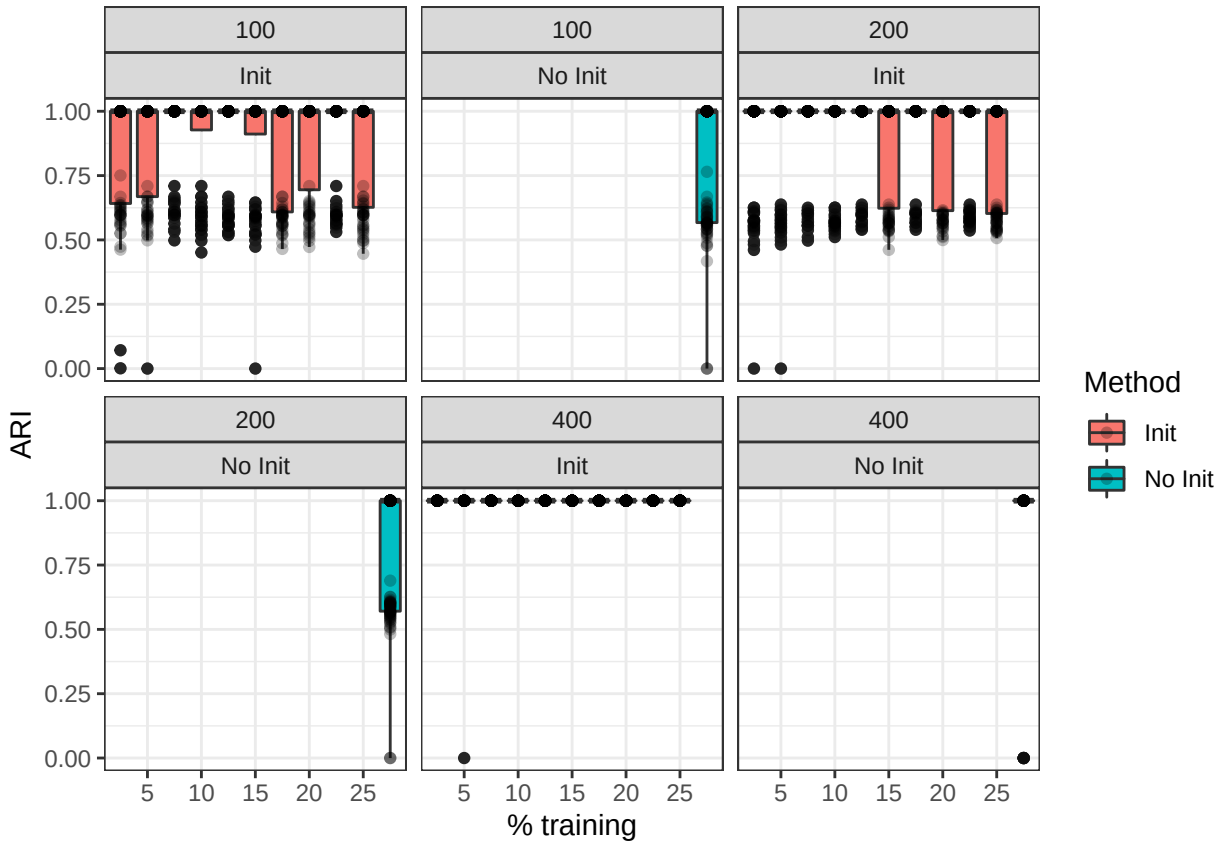
Dense Setting

This is a quite easy simulation setting but the initialization procedure still leads to a slight improvement. Here `no_init` corresponds to the random initialization, which shows a lot more variance.

```
## [1] "Rate Matrix"
```

```
##      [,1] [,2]
## [1,] 0.50 0.05
## [2,] 0.05 1.00
```

```
## # A tibble: 11 x 4
## # Groups:   init [2]
##    init       n0 mean_ARI sd_ARI
##    <chr>   <dbl>    <dbl>  <dbl>
##  1 Init        5    0.906  0.195
##  2 Init       10    0.906  0.195
##  3 Init       15    0.943  0.143
##  4 Init       20    0.919  0.168
##  5 Init       25    0.934  0.153
##  6 Init       30    0.909  0.183
##  7 Init       35    0.920  0.166
##  8 Init       40    0.912  0.173
##  9 Init       45    0.928  0.157
## 10 Init       50    0.889  0.188
## 11 No Init    NA    0.814  0.263
```



We also consider a setting with $K = 3$ which is harder for the random starting point to recover the communities.

```
K <- 3
true_Mu <- matrix(0.05,
                  nrow = K, ncol = K, byrow = T)
diag(true_Mu) <- 0.5:K + 1.2
print(true_Mu)
```

```
##      [,1] [,2] [,3]
## [1,] 1.70 0.05 0.05
## [2,] 0.05 2.70 0.05
## [3,] 0.05 0.05 3.70
```

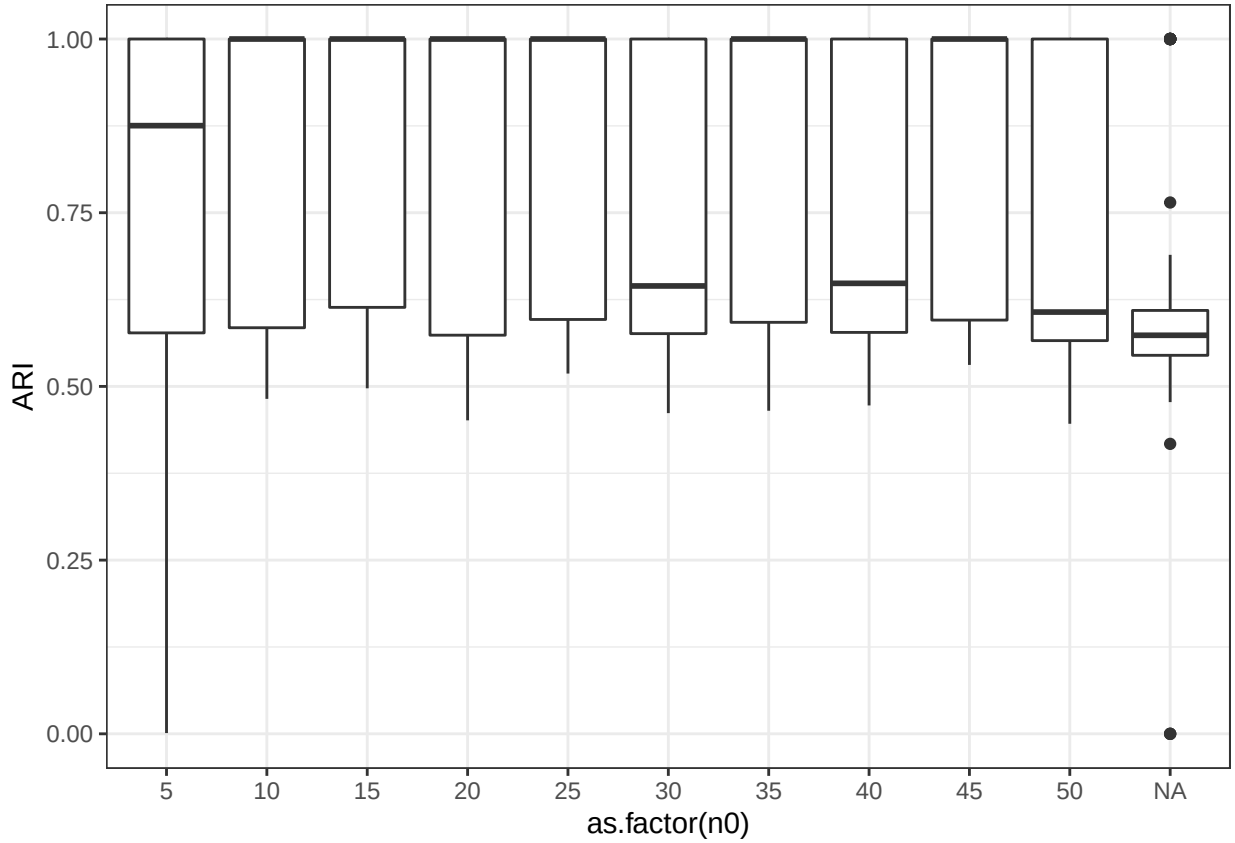
```
exp_12_sim_2 <- list.files(path = here("Experiments/exp_results/"),
                           pattern = "exp_12_n0_dense_apr_13")

exp_12_sim_2 %>%
  map_dfr(~readRDS(here("Experiments/exp_results/", .x))) %>%
  group_by(n0) %>%
  summarise(mean(ARI), sd(ARI))
```

```
## # A tibble: 11 x 3
##       n0 `mean(ARI)` `sd(ARI)`
##    <dbl>       <dbl>     <dbl>
```

```
##  1     5       0.784     0.228
##  2    10       0.795     0.213
##  3    15       0.857     0.198
##  4    20       0.797     0.215
##  5    25       0.834     0.206
##  6    30       0.781     0.217
##  7    35       0.800     0.212
##  8    40       0.780     0.215
##  9    45       0.820     0.206
## 10    50       0.722     0.204
## 11    NA       0.605     0.182
```

```
exp_12_sim_2 %>%
  map_dfr(~readRDS(here("Experiments/exp_results/", .x))) %>%
  ggplot(aes(as.factor(n0), ARI)) +
  geom_boxplot()
```



Sparse Setting

Here $\rho = 0.1, 0.05, 0.02$ and `no_init` corresponds to the random initialization, which shows a lot more variance and doesn't recover the clusters as well.

Here for the small sparse network our initialization procedure still struggles but as the number of nodes increases we can reliably recover the network structure well.

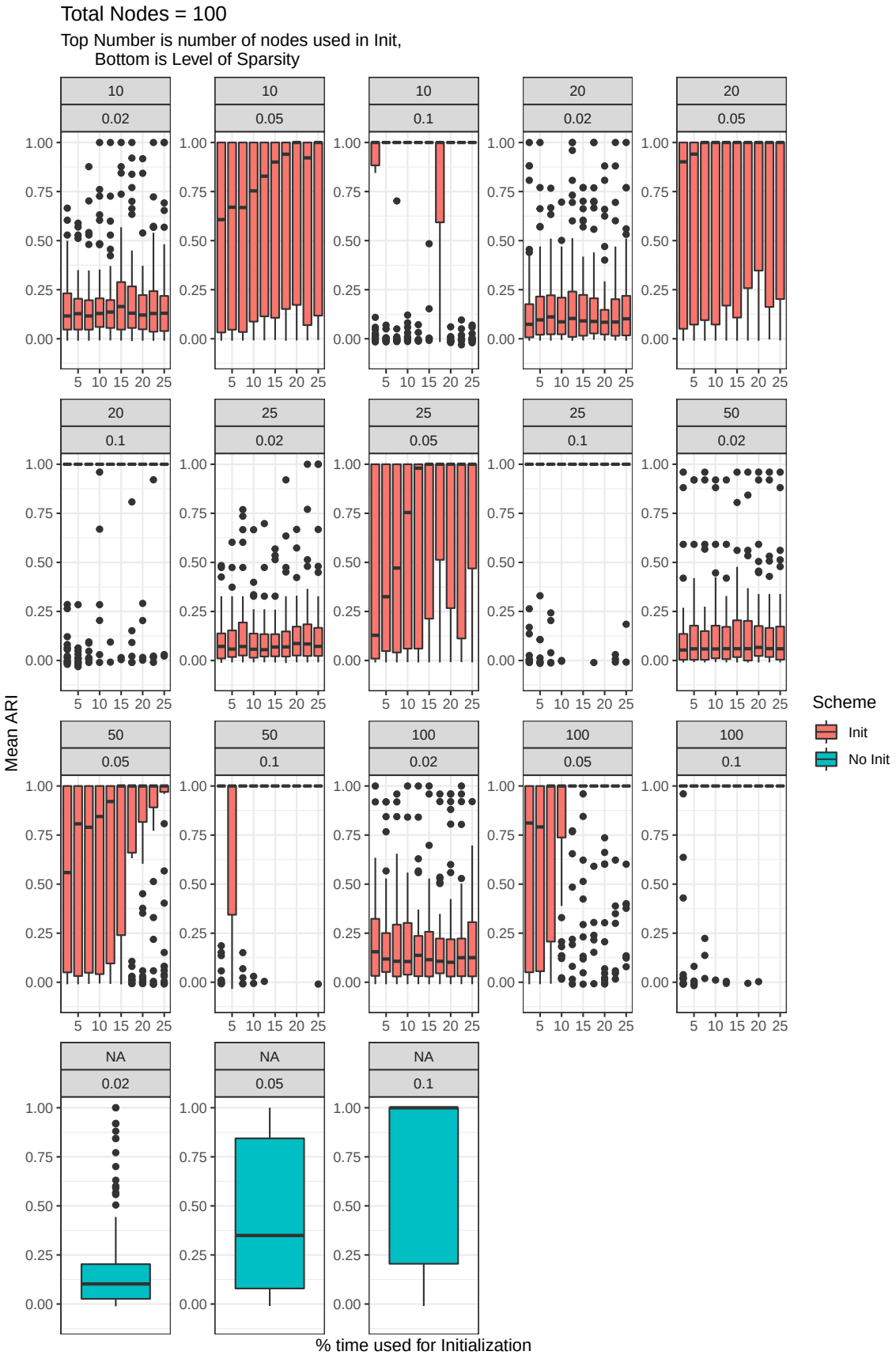
```
## Warning in matrix(c(2, 0.05, 0.4, 1.5), nrow = K, ncol = K, byrow = T): data
## length [4] is not a sub-multiple or multiple of the number of rows [3]
```

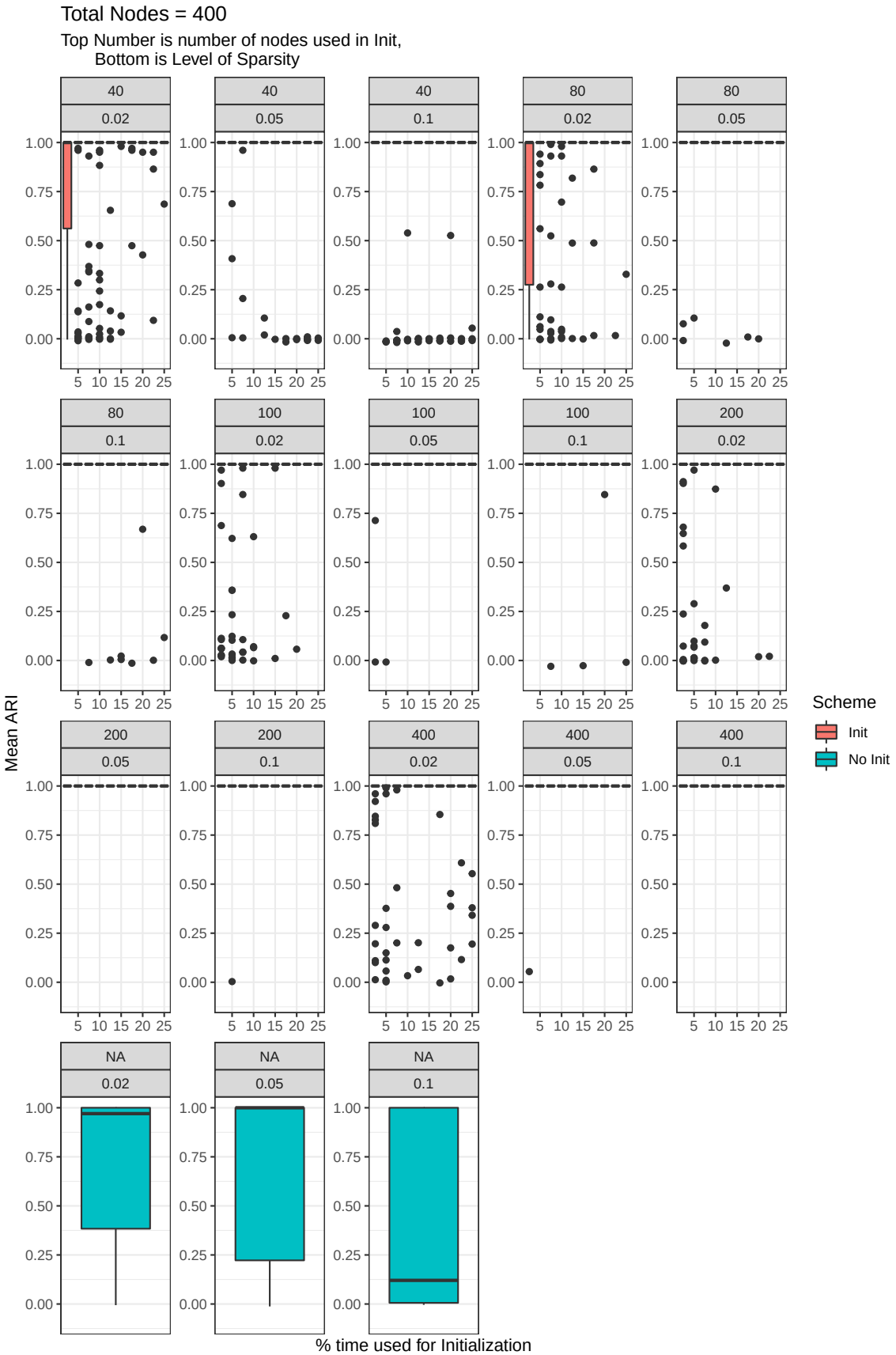
```
## [1] "Rate Matrix"
```

```
##      [,1] [,2] [,3]
## [1,]  2.0 0.05 0.40
## [2,]  1.5 2.00 0.05
## [3,]  0.4 1.50 2.00
```

```
## `summarise()` has grouped output by 'init'. You can override using the `.groups` argument.
```

```
## # A tibble: 11 x 4
## # Groups:   init [2]
##    init       n0 mean_ARI sd_ARI
##    <chr>   <dbl>    <dbl>  <dbl>
##  1 Init        5    0.712  0.417
##  2 Init       10    0.731  0.407
##  3 Init       15    0.762  0.387
##  4 Init       20    0.786  0.371
##  5 Init       25    0.802  0.362
##  6 Init       30    0.814  0.354
##  7 Init       35    0.812  0.356
##  8 Init       40    0.818  0.350
##  9 Init       45    0.817  0.353
## 10 Init       50    0.819  0.351
## 11 No Init    NA    0.541  0.424
```


Total Nodes = 100
Top Number is number of nodes used in Init,
Bottom is Level of Sparsity
10
0.02
10
0.05
10
0.1
20
0.02
20
0.05
20
0.1
25
0.02
25
0.05
25
0.1
50
0.02
50
0.05
50
0.1
100
0.02
100
0.05
100
0.1
NA
0.02
NA
0.05
NA
0.1
Mean ARI
% time used for Initialization
Scheme
Init
No Init


Total Nodes = 400
Top Number is number of nodes used in Init,
Bottom is Level of Sparsity
40
0.02
40
0.05
40
0.1
80
0.02
80
0.05
80
0.1
100
0.02
100
0.05
100
0.1
200
0.02
200
0.05
200
0.1
400
0.02
400
0.05
400
0.1
NA
0.02
NA
0.05
NA
0.1
Mean ARI
% time used for Initialization
Scheme
Init
No Init

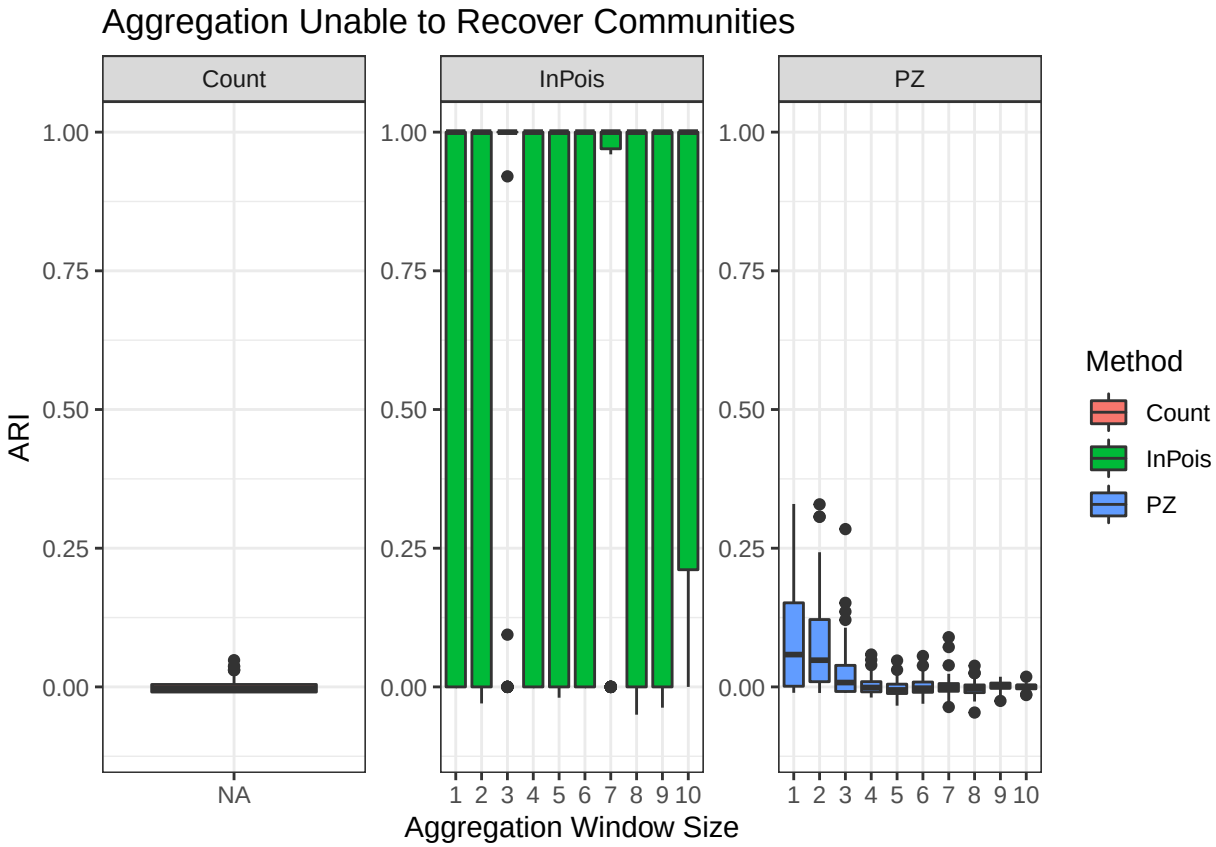
Experiment for Figure 1

```
fig1 <- list.files(path = here("Experiments/exp_results/"),
                           pattern = "fig_1_exp_1_apr_13")

fig1 %>%
  map_dfr(~readRDS(here("Experiments/exp_results/", .x))) %>%
  mutate(window_size = as.factor(window_size)) %>%
  drop_na(ARI) %>%
  group_by(Method) %>%
  summarise(mean_ari = mean(ARI), sd_ari = sd(ARI))
```

```
## # A tibble: 3 x 3
##   Method mean_ari sd_ari
##   <chr>     <dbl>  <dbl>
## 1 Count   0.00186 0.0149
## 2 InPois  0.675   0.467
## 3 PZ      0.0211  0.0580
```

```
fig1 %>%
  map_dfr(~readRDS(here("Experiments/exp_results/", .x))) %>%
  mutate(window_size = as.factor(window_size)) %>%
  drop_na(ARI) %>%
  ggplot(aes(window_size, ARI, fill = Method)) +
  geom_boxplot() +
  facet_wrap(~Method, scales = "free") +
  ylim(c(-0.1, 1)) +
  labs(x = "Aggregation Window Size",
       title = "Aggregation Unable to Recover Communities")
```


Aggregation Unable to Recover Communities
Count
InPois
PZ
ARI
1.00
0.75
0.50
0.25
0.00
NA
1 2 3 4 5 6 7 8 9 10
Aggregation Window Size
Method
Count
InPois
PZ