Sank Freaking Sanity

Travis Weber

2022-08-03

Include statements

```
library(reshape)
library(tidyr)
library(ggplot2)
library(ggalluvial)
library(alluvial)
library(plyr)
library(dplyr)
library(cluster)
source("./04_analysis/02_protocol/Utils.r")
```

Data import

```
source("./04_analysis/01_input/setup.r")
binaryAttrs <- read.csv("./04_analysis/01_input/BinaryAttrs.csv") %>%
  dplyr::rename(filename = binary)
```

Clustering

```
clust_and_title <- function(col) {
  clust <- pam(col, 3)
  clust_titles(clust)
}
```

Cluster By Version

Cluster by size of binary

```
binaryAttrs <- binaryAttrs %>% mutate(size_cluster = clust_and_title(size))
```

Visualizing clusters / attributes

Step 1. Add cluster data to long format

```
cve_bin_long_clusts <- left_join(cve_bin_long, cve_version_clusters)
```

Step 2. Add Binary attribute to long format

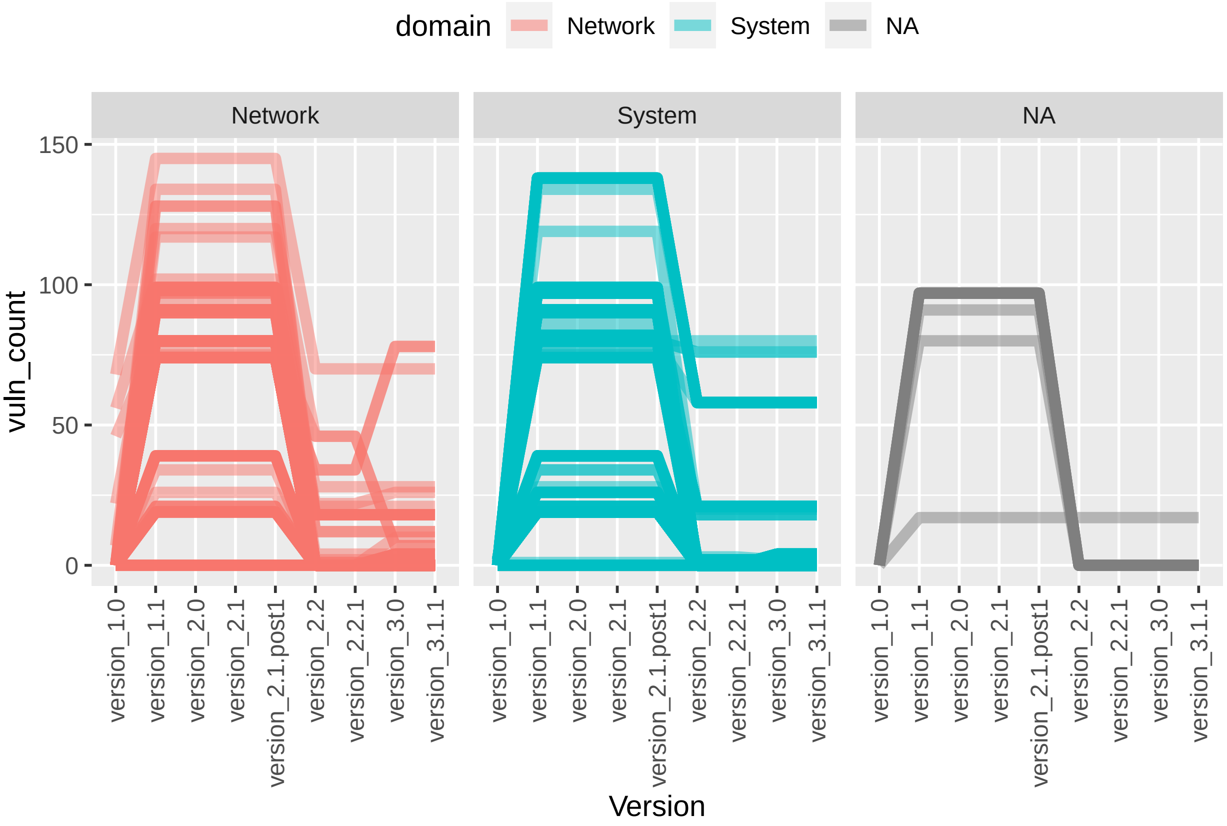
```
cve_bin_long_clusts <- left_join(cve_bin_long_clusts, binaryAttrs)
```

Step 3. Define a graphing function

```
split_graph_by_factor <- function(data, fact) {
  ggplot( mapping = aes(x = version, y = vuln_count)) +
  geom_line(mapping = aes(group = filename, color = fact), size = 2, alpha = 0.5) +
  scale_x_discrete(guide = guide_axis(angle = 90)) +
  xlab("Version") + ylab("Findings Count") +
  facet_grid(cols = vars(fact)) +
  theme(legend.position = "none") +
  scale_fill_brewer(type = "qual")
}
```

Step 4. Make some graphs

```
cve_bin_long_clusts %>% ggplot( mapping = aes(x = version, y = vuln_count)) +
  geom_line(mapping = aes(group = filename, color = domain), size = 2, alpha = 0.5) +
  scale_x_discrete(guide = guide_axis(angle = 90)) +
  xlab("Version") +
  facet_grid(cols = vars(domain)) +
  theme(legend.position = "top") +
  scale_fill_brewer(type = "qual")
```



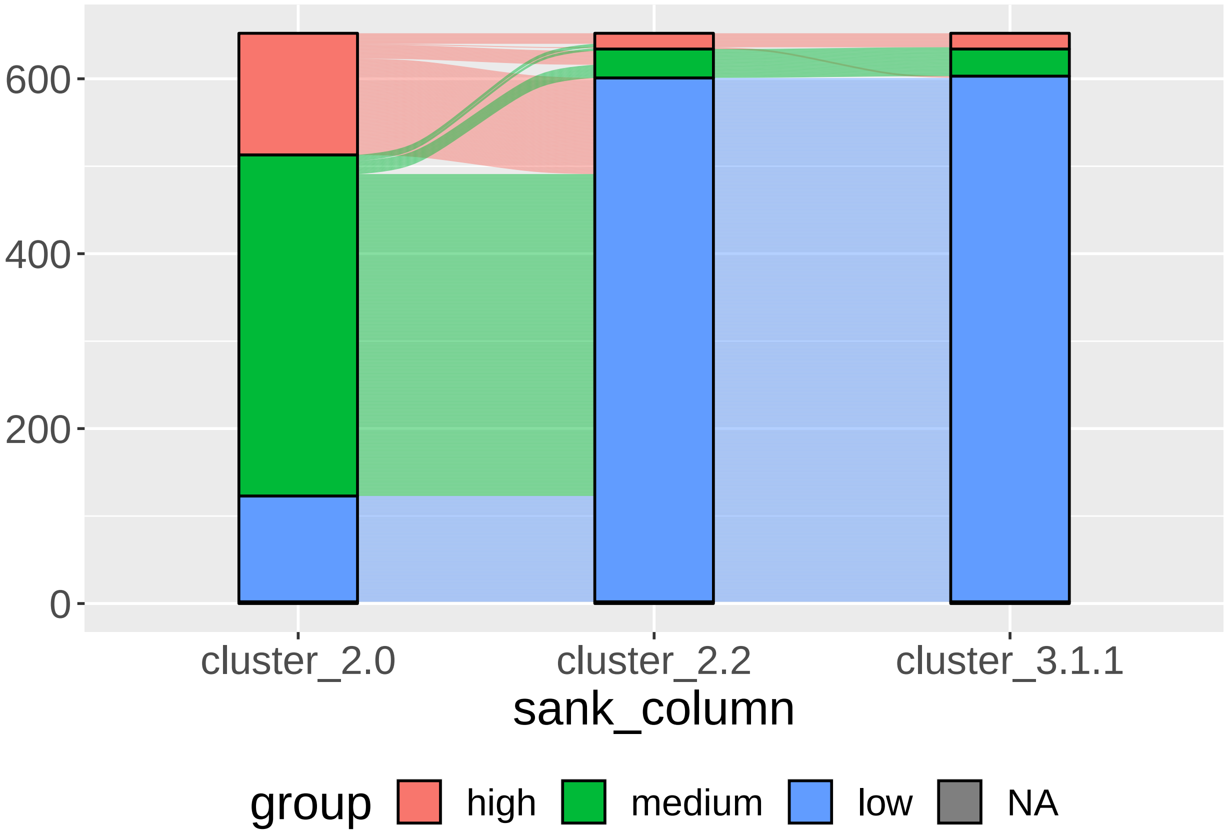
Sankey's

sankification function

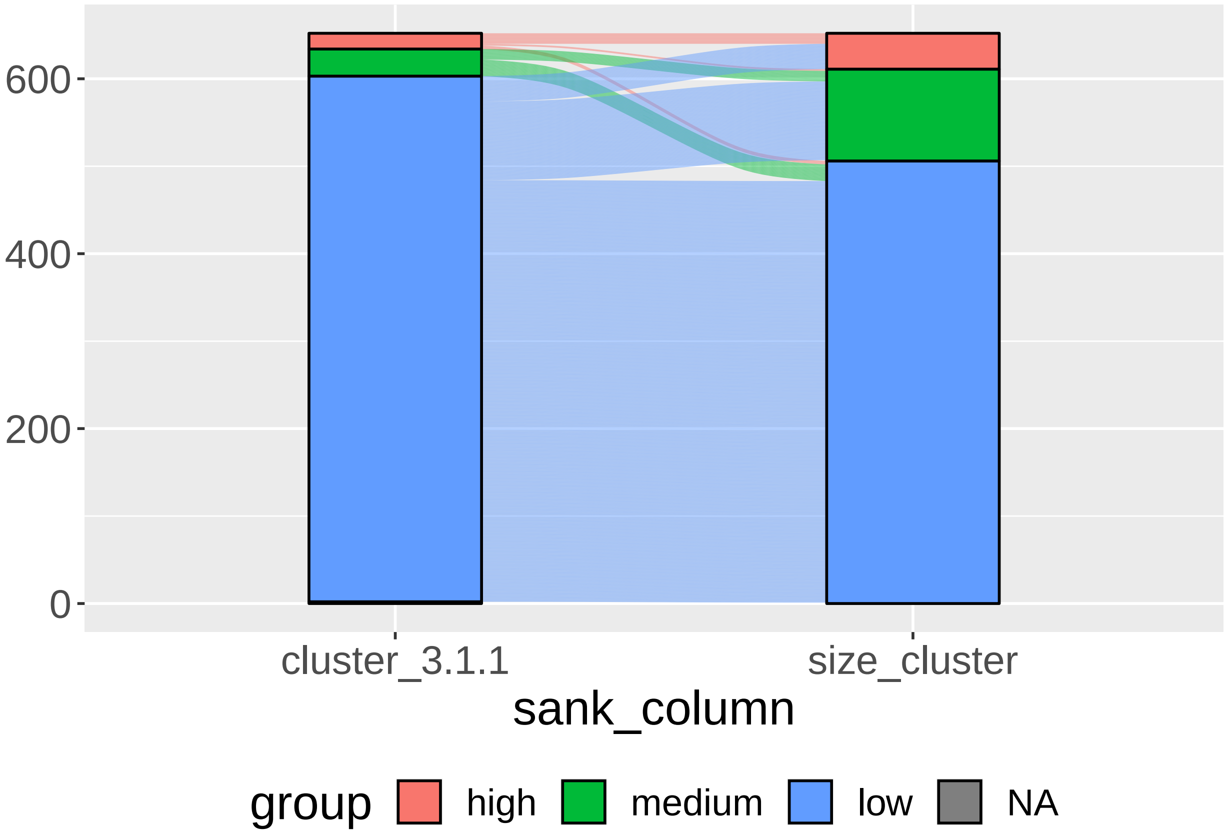
```
sankey <- function(columns) {
  columns <- c(columns, "filename")

  version_clusters_wanted <- cve_version_clusters %>%
    select(any_of(columns))

  binary_attributes_wanted <- binaryAttrs %>%
    select(!size) %>%
    select(any_of(columns)) %>%
    mutate_all(as.factor)

  binary_attributes_wanted %>%
    left_join(version_clusters_wanted) %>%
    pivot_longer(
      cols = !filename,
      names_to = "sank_column",
      values_to = "group"
    ) %>%
    mutate(sank_column = factor(sank_column, levels = columns)) %>%
    ggplot(aes(x = sank_column, stratum = group, alluvium = filename,
           fill = group, label = group)) +
    geom_flow(stat = "alluvium", lode.guidance = "frontback") +
    geom_stratum() +
    theme(
      legend.position = "bottom",
      text = element_text(size = 18),
      axis.text = element_text(size = 15)
    )
}
```

Make graphs

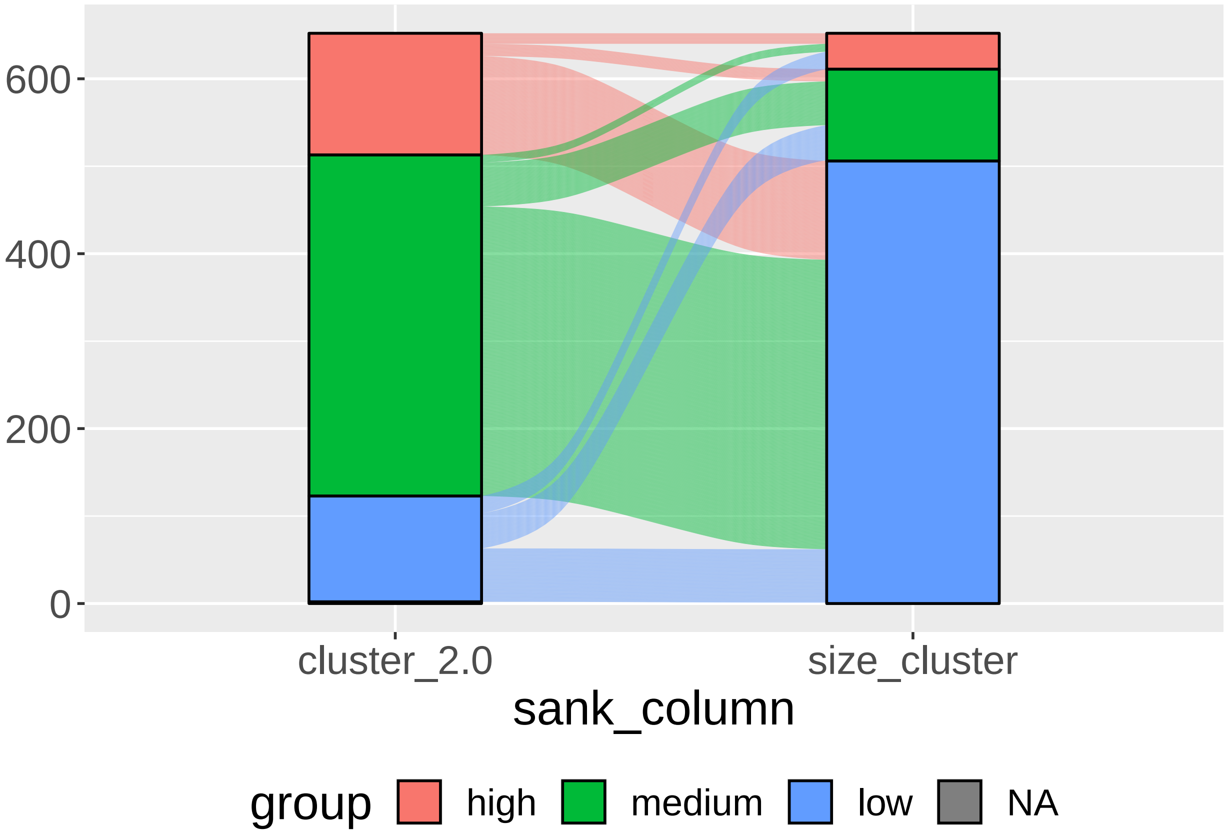
```
sankey(c("cluster_2.0", "cluster_2.2", "cluster_3.1.1"))
```



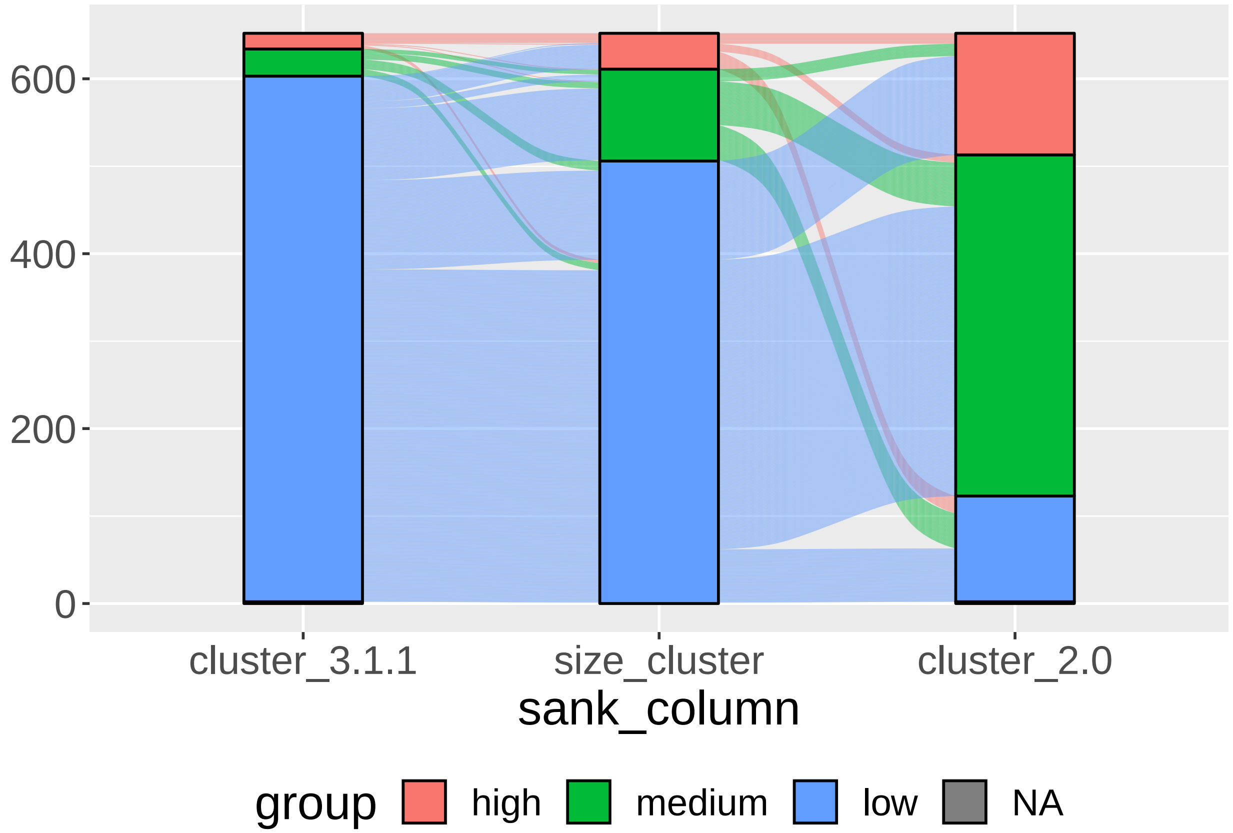
```
sankey(c("cluster_3.1.1", "size_cluster"))
```



```
sankey(c("cluster_2.0", "size_cluster"))
```



```
sankey(c("cluster_3.1.1", "size_cluster", "cluster_2.0"))
```


600
400
200
0
cluster_3.1.1
size_cluster
cluster_2.0
sank_column
group
high
medium
low
NA