Analysis of Topic Networks on Stack Overflow

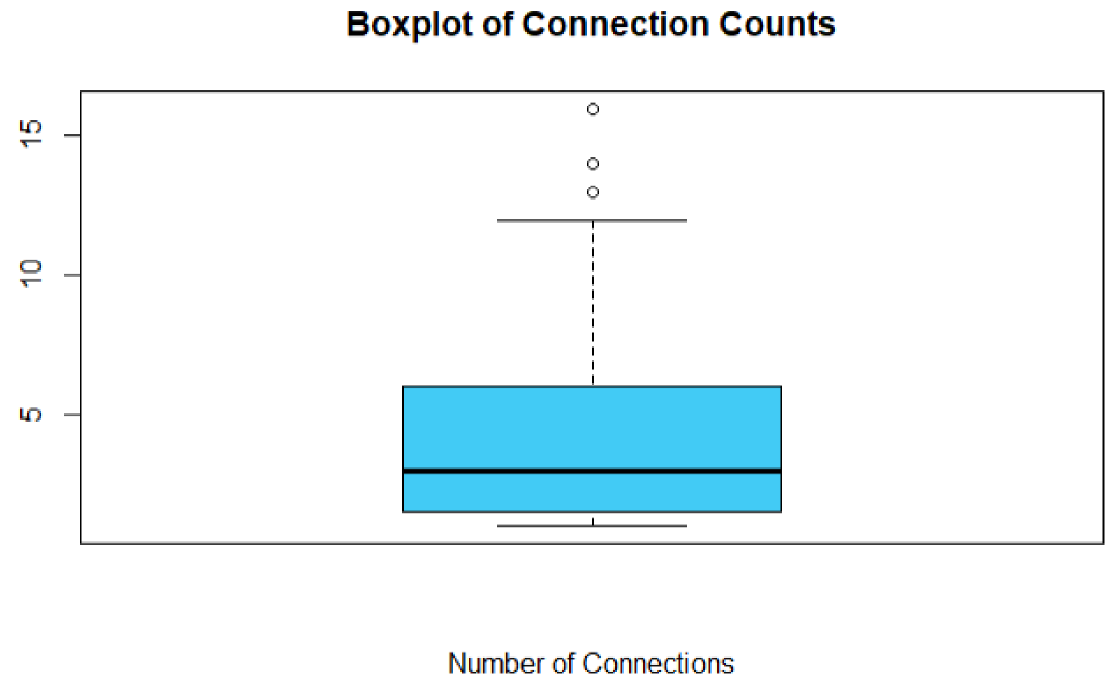
November 8th, 2023

Report Presented by:

Carson Edwards

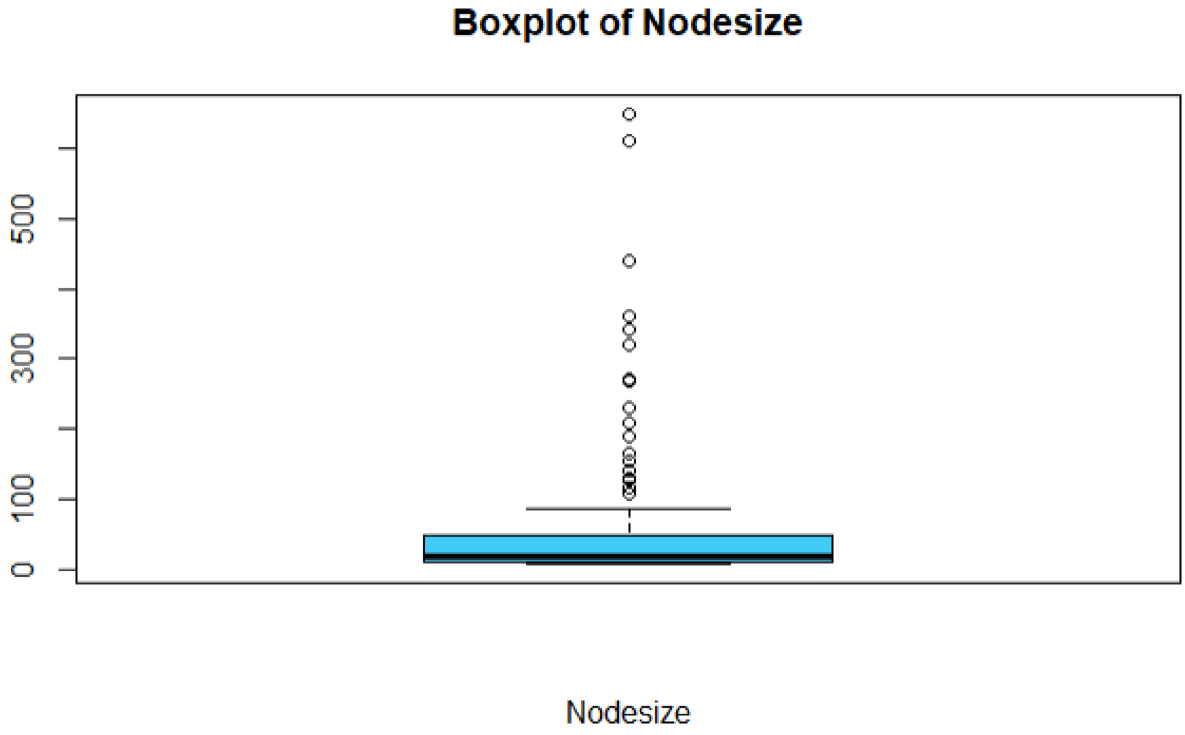
Stack Overflow is one of the most popular websites for programmers to get questions answered by other programmers. It was created in 2008 but as of March 2022 has over 20 million registered users (All Sites - Stack Exchange). This analysis focuses on a dataset from Kaggle that contains network information about tag data on posts (Stack Overflow Tag Network). The goal is to understand how tags can be related to one another. Through this analysis one could get an idea of tags they are not familiar with based on how the tags connect to well known tags such as "python".

We will start by getting an idea of the distribution of the number of connections. We will take a look at this with a box plot showing connection counts.

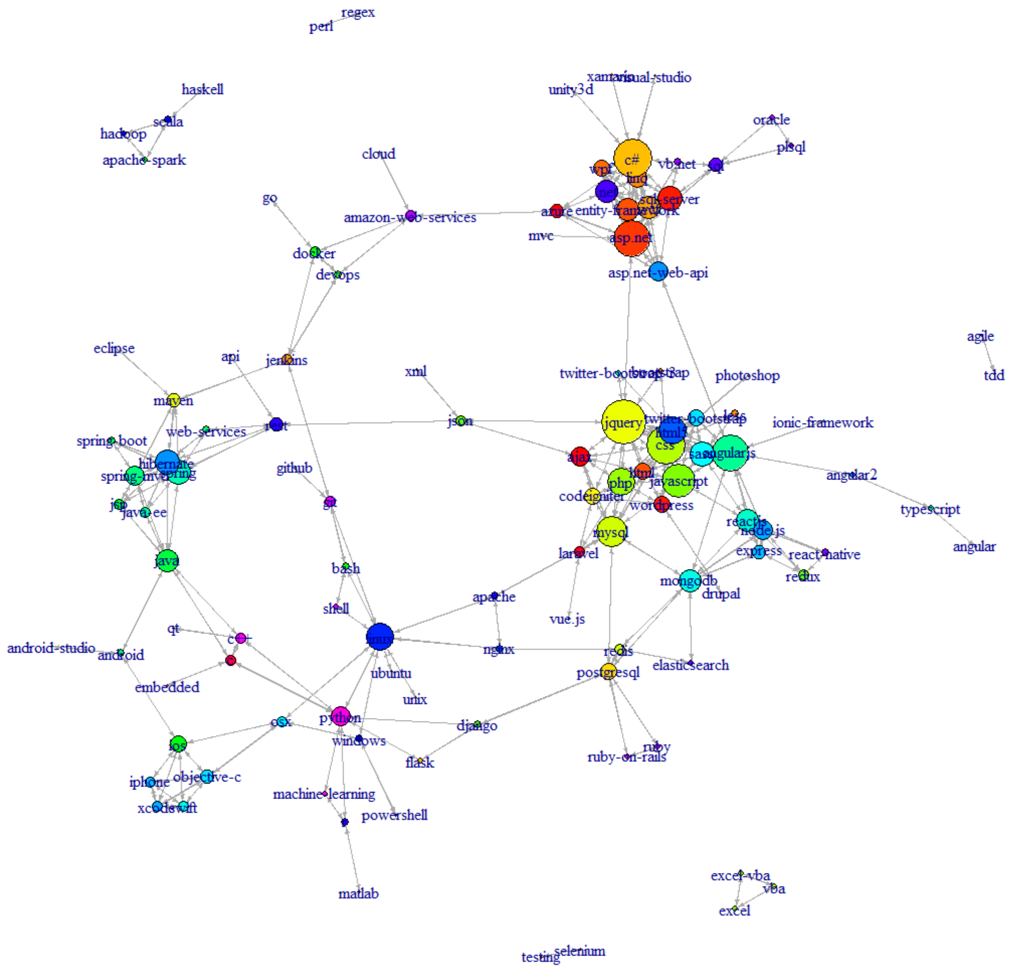


The box plot tells us that most of the tags only connect with 5 or less other tags. However, there are some tags with upwards of 10 to 15 connections. These tags will likely be the ones that create distinct communities. From an educational learning perspective, a student who is interested in growing their skills is going to most likely be familiar with the tags with the most connections. Students could use this information to explore other topics that would help their natural learning progression.

Below is a boxplot showing node size from our nodes table. This variable is a representation of how often certain tags are used. This chart is interesting because it shows how many outliers are in our data and that most tags are used far less often than these outliers. We can keep the node size in mind as we evaluate our network graphs, since tags with smaller node sizes may not show complete connections to all the tags they might actually be associated with in practice.

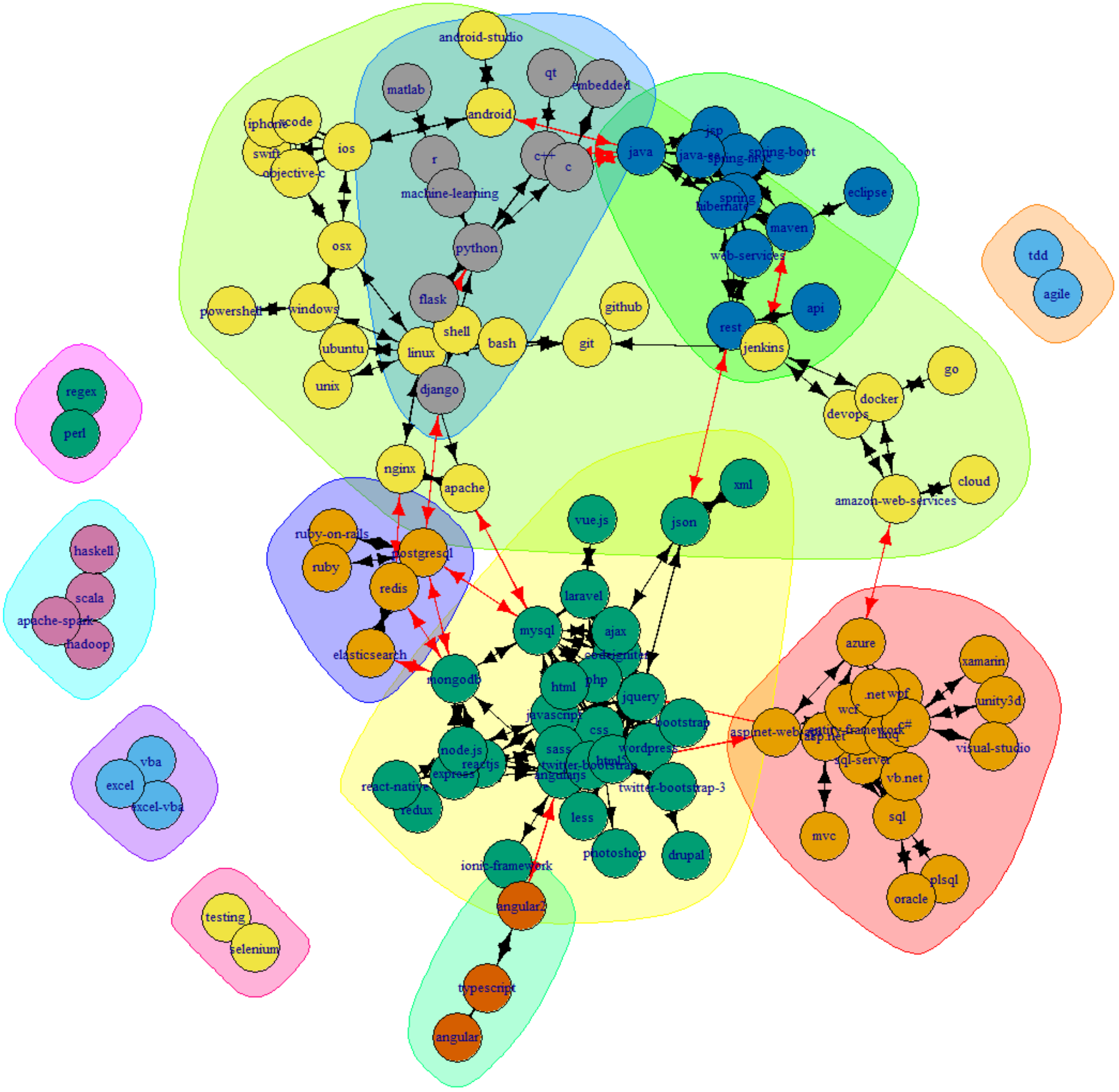


Next, we will take a look at a network graph of these connections. The layout chosen for this graph is based on the Fruchterman-Reingold algorithm. This layout does a good job of spacing out our nodes and this helps with overall readability (Reingold Layout). It also presents us with a vague idea of what our communities may be within this data.



From the network graph above, there are a few key things that we can immediately pick up on. As expected, there are high connection tags that are well known concepts such as jquery, css, mysql, linux, c#, java, and others. These have connections to some other less known topics such as mysql connecting with laravel and codeigniter. It is worth noting that my personal experience plays a significant role in identifying which of these topics are more “well known” than others. However, I do think that evaluating nodes with this idea of popularity based on number of connections makes sense. These popular topics will play a role in identifying key features of our communities within the data.

Below is a network graph highlighting densely connected nodes. This gives us an idea of the communities of tags in Stack Overflow.



There are a few interesting conclusions we can gather from this graph. The first is that many of our communities overlap tags, but some do not. For example, perl and regex are grouped in their own community, which is interesting because regex is not a concept only associated with Perl. When evaluating node size, we can see that regex has a node size of 9.46

and perl has a node size of 19.38. These values are very low compared to others in the data set, so it is likely that these tags are underrepresented and we cannot get a full idea of how connected they are to all the other tags. Other small communities make sense such as angular, angular2, and typescript all being in a group together.

The overlapping communities show interesting insight about how related topics are. The angular and typescript community is connected to the javascript community on the angularjs tag. There is a general operating system development community shown in the light green containing linux, ubuntu, android, and windows which is surrounding the python/c++/machine learning community. A student or professional interested in expanding their skill sets could use this graph to see natural stepping stones in learning. They could start with any subject such as html and then expand into javascript, css, or reactjs. Becoming comfortable with that community of subjects would allow them to branch out from any direction toward other communities. For example, once they become familiar with mysql and mongodb they could move into learning about elasticsearch and ruby-on-rails.

In conclusion, this analysis of tag data provides valuable information for someone who is interested in learning new skills. Viewing the network graph with communities easily allows for an understanding of how connected certain topics are. This can also be used to understand terms one may not be familiar with, by assessing how connected the unknown term is to a term one does know.

References

All Sites - Stack Exchange. (n.d.). Stackexchange.com. Retrieved November 8, 2023, from https://stackexchange.com/sites?view=list#users

finnstats. (2021, April 22). *Social Network Analysis in R | R-bloggers*. https://www.r-bloggers.com/2021/04/social-network-analysis-in-r/

Reingold Layout - an overview | ScienceDirect Topics. (n.d.). Www.sciencedirect.com. https://www.sciencedirect.com/topics/computer-science/reingold-layout

Stack Overflow Tag Network. (n.d.). Www.kaggle.com. Retrieved November 8, 2023, from https://www.kaggle.com/datasets/stackoverflow/stack-overflow-tag-network/