

For this assignment, I chose the BigDataatwitter.graphml file which is available at ([https://docs.google.com/file/d/0B385K\\_-eDQPxenKzZzdZUDM0RlE/edit?pli=1](https://docs.google.com/file/d/0B385K_-eDQPxenKzZzdZUDM0RlE/edit?pli=1)). This dataset is composed of twitter data of the BigData followers.

I first ran the contraction algorithm followed by Force Atlas algorithm. Then I ran the Noverlap algorithm to avoid the overlapping of nodes while keeping the shape of the graph. To configure node's color and size, I selected "Degree" as a ranking parameter. The size of nodes vary from 30 to 200. I then ran the Average degree statistics and Modularity (using weights of edges) to identify communities in the given graph. Label adjust algorithm has also been used to avoid the overlapping of node labels. Since this was a large graph, I filtered the nodes with less than 5 degree.

As mentioned earlier, I have portioned the nodes by modularity. Each node belongs to a separate community or a cluster. And each community or cluster has been colored accordingly, making each community distinct from the other. As we can see from the graph, there are six different communities in the given graph, each denoted by light green, dark green, maroon, blue, dark pink and brown color codes.

We can make the following observations from the graph:

- "biconnections" and "hortonworks" are the two most important Hadoop communities.
- The light green community is composed of twitter handles who often tweet about the "BIconnections", which stands for Business Intelligence connections.
- The maroon community comprises of twitter handles who most oftenly tweet or retweet about Horton works.
- The dark green community comprises of twitter users who tweet about both BIconnections and Hortonworks and most probably follow both the communities.
- The brown community also comprises of users who have more preference to Hortonworks but tweet less often about HortonWorks.
- The dark pink community consists of twitter users who used to tweet about hortonworks but now tweet less about it.

Other observations:

- Hortonworks has the highest degree followed by BIconnections.
- BIconnections has stronger edge weight to eric\_kavanagh twitter handle. This could be because eric\_kavanagh very oftenly tweets about BIconnections.
- hsteph node serves the role of a bridge. It is one of the most important node in the graph as it connects three different communities indicated by maroon, green and dark pink color codes.
- BIconnections, hortonworks, merv, mjasay, jimplush are hubs.

Below is the final graph obtained from the dataset.

