Vladimir Nimchenko
Evan McLaughlin

We elected to work on a Wikipedia data extraction script.

The code extracts data from Wikipedia, focusing on four sports categories: Ice Hockey, American Football, Baseball, and Basketball. It constructs a network graph where pages are nodes, and connections are edges based on category membership. The script then calculates degree centrality and eigenvector centrality for each page within the network, offering insights into the connectivity and importance of pages within each sports category.

First up, the degree centrality values, which represent the number of connections each page has within its category. The degree centrality numbers suggest that Ice Hockey has the highest connectivity, followed by American Football, Baseball, and Basketball. This implies that, within the Wikipedia pages of these sports categories, Ice Hockey has more interconnected pages compared to others.

Now, moving on to eigenvector centrality, a measure considering both direct and indirect connections. The eigenvector centrality results paint a slightly different picture. Here, American Football takes the lead, indicating that pages within the American Football category not only have numerous connections but also connect with other well-connected pages. Ice Hockey, Baseball, and Basketball follow, but the relative positions have shifted compared to degree centrality. This suggests that the importance of a page in American Football is influenced not just by its connections but also by the connections of the pages it's connected to.

Lastly, we ran a t-test to assess the statistical significance of the differences between the two centrality measures. The t-test results we generated were remarkably high at 21.21 and a very low p-value of 1.18e-26. These results indicate a statistically significant difference between the degree centrality and eigenvector centrality values for the given page categories. This suggests a substantial disparity in connectivity patterns, emphasizing the distinct roles and structures captured by these two centrality measures in the network.