**I. Introduction.** Describe the data source, the analytic question assigned, and how you chose to address that question.

Our data source for this research assignment was last.fm. We wanted to compare the 60s and 70s tags and the networks that would arise from them. When we first went about doing this, we noticed that the top artist in the 60s tag, The Beatles, were very central to the network. Comparatively, the top artist in the 70s tag, Led Zeppelin, was not very central in their network. A lot of the tags that The Beatles were connected to were also connected to other artists in the 60s, but Led Zeppelin had tags that were different from the other artists connected to the 70s tag. This difference interested us. We wondered if this distinction would maintain with more nodes in the network. Our hypothesis was that the 60s network would have a higher connectivity, and that the 70s network would be more divided into specific communities.

In order to address this question, we wanted to get the top artists from the 60s and 70s tags and then find out what other tags those artist are connected to. We wanted to see if either of the networks could be broken into smaller communities. We also wanted to calculate the average centralities of the network. If we found that the 60s graph was more centralized and the 70s graph had more distinct communities, our hypothesis would be proven correct.

**II. Data Gathering.** Describe the data gathering methodology. Note any difficulties presented by the source or the data. (For example, rate limits.) What entities were chosen as nodes and edges?

Initialized the network with 41 nodes - the decade tag (either ‘60s’ or ‘70s’) and the top 40 artists associated with the tag. We had no problems with being rate limited. If the rate limits exist, they are perhaps applied automatically. We have created two lastfm accounts just in case. “Tag” objects returned by pylast were converted to strings, as were “Artist” objects. Tags were bipartite==0, artists were bipartite==1

Every Artist returned (up to) the top 10 tags associated with them, and every Tag returned (up to) the top 10 artists. These numbers were chosen arbitrarily. Depending on the tag there are potentially thousands of possible results, and as any user can create a tag each artist can potentially have many tags as well. 10 seemed like a good amount. Depth crawled was to 5, nodes crawled were 2000.

*Initial 60s artists were:*

['The Zombies', 'The Monkees', 'The Hollies', 'The Turtles', 'Small Faces', 'The Ronettes', "The Lovin' Spoonful", "Herman's Hermits", 'The Shangri-Las', 'Del Shannon', 'The Crystals', 'The Shirelles', 'Tommy James & The Shondells', 'Petula Clark', 'The Spencer Davis Group', 'The Box Tops', 'Lesley Gore', 'The Searchers', 'The 5th Dimension', 'The Archies', 'The Walker Brothers', 'Gerry & The Pacemakers', 'Sonny & Cher', 'Skeeter Davis', 'The Association', 'The Easybeats', 'The Chiffons', 'Mama Cass', 'Nancy Sinatra & Lee Hazlewood', 'Gene Pitney', 'The Tremeloes', 'Paul Revere & The Raiders', 'Dave Dee, Dozy, Beaky, Mick & Tich', 'Sandie Shaw', 'Jackie DeShannon', 'The Dixie Cups', 'Little Eva', 'The Dave Clark Five', 'The Beau Brummels', 'Sam The Sham & The Pharaohs']

*Initial 70s artists were:*

['Albert Hammond', 'Bay City Rollers', 'Looking Glass', 'Lobo', 'Player', 'David Essex', 'The Rubettes', 'Andrew Gold', 'Middle Of The Road', 'Showaddywaddy', 'The New Seekers', 'Sailor', 'Tony Orlando & Dawn', 'Harpo', 'Rita Lee & Tutti Frutti', 'Lips', 'Les Humphries Singers', 'Hamilton, Joe Frank & Reynolds', 'David Soul', 'LoCash', 'Racey', 'Walter Egan', 'Clout', 'Dave Loggins', 'Samantha Sang', 'Daniel Boone', 'Spiral Starecase', 'Elia y Elizabeth', 'Morris Albert', 'Chicory Tip', 'Johnny Wakelin', 'Peter Allen', 'Chris Norman & Suzi Quatro', 'Leif Garrett', 'The Wombles', 'Barry Blue', 'Arthur Fiedler & The Boston Pops Orchestra (Holiday)', 'Tony Orlando', 'Kincade', 'Pholhas']

**III. Data.** Describe the data gathered: the number of nodes of each bipartite type and types and ranges of attributes. Describe any filtering and projection operations performed on the data before analysis. For each final network, indicate the density, diameter, and include visualizations of the degree Distributions.

In terms of the actual data, we filtered it by removing erroneous connections between nodes of bipartite 0 and bipartite 1. We also removed self-loop edges, and confirmed that the graph was fully bipartite. After doing calculations for many of the attributes listed in the tables below, including categorizing each entry into a specific community, we did weighted projections for artist-to-artist and tag-to-tag graphs.

We found it necessary to filter our data in a few ways. There were of course some nodes that had too few connections for them to be useful to us, so we filtered out nodes which had degree of 6 or less. There were also some tag nodes which had a lot of connections, but did not tell us very much. An example of this was the ‘seen live’ tag. We did not need tags like this because almost every artist node was connected to it, so it did not give us any new information. Because of this, we also filtered out for nodes which had too many connections. Another way we filtered data is by the attribute “Distance from (60s/70s) Node”, which basically filtered by distance from the beginning node.

**Data:**

\*Note: Clustering for the bipartite graphs is robins\_alexander\_clustering while for the projected graphs it is an approximation of clustering

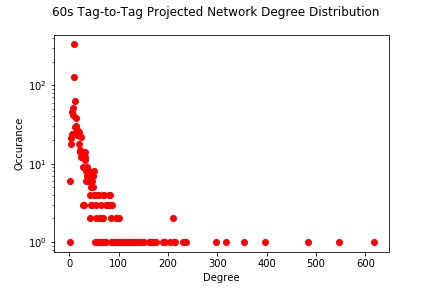
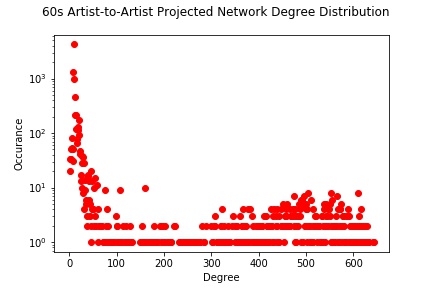
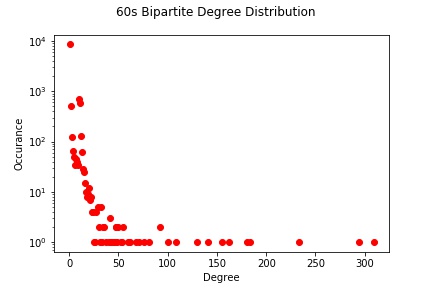
Number of communities is found using the new greedy\_modularity\_communities in networkx 2.2

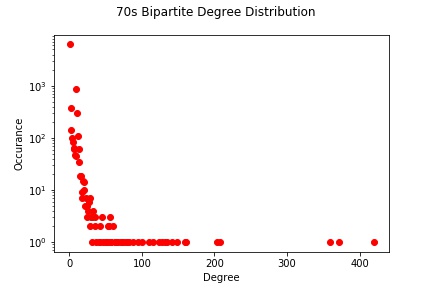
Density for bipartite graphs is in nx.algorithms.bipartite.density while density for the projected graphs is just nx.density

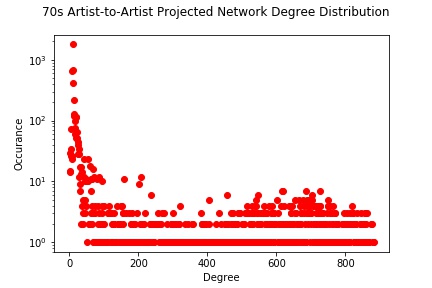
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Size | Density | Diameter | Clustering Coefficent | Number of communities |
| 60s Bipartite Network | 11098 | 0.0012726597717024326 | 10 | 0.1701168439251565 | 306 |
| 60s Artist-to-Artist Projected Network | 9745 | 0.004399692149820421 | 5 | 0.95 | 427 |
| 60s Tag-to-Tag Projected Network | 1353 | 0.018138521890867107 | 4 | 0.736 | 9 |
| 70s Bipartite Network | 8739 | 0.0012000234558478477 | 10 | 0.13526917458044976 | 194 |
| 70s Artist-to-Artist Projected Network | 6508 | 0.01536050864422967 | 5 | 0.91 | 272 |
| 70s Tag-to-Tag Projected Network | 2231 | 0.009861048857014791 | 4 | 0.809 | 32 |

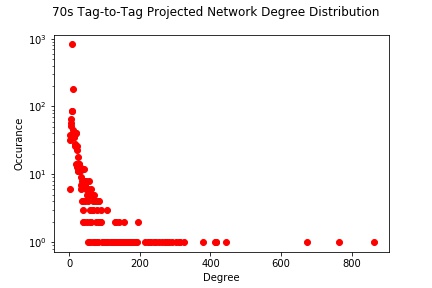
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attributes PART1 | BipartiteClusteringCoefficient | BipartiteClosenessCentrality | BipartiteDegreeCentrality | BipartiteBetweennessCentrality |
| 60s Bipartite | 0.00913 - 0.95 | 0.16037 - 0.46042 | 0.000103 - 0.031709 | 0.0 - 0.234748 |
| 60s A2AP | 0.0319 - 0.95 | 0.23491 - 0.460422 | 0.000739 - 0.023651 | 0.0 - 0.018604 |
| 60s T2TP | 0.009133 - 0.683333 | 0.160378 - 0.292402 | 0.000103 - 0.031709 | 0.0 - 0.23474824 |
| 70s Bipartite | 0.007999 - 0.925926 | 0.181625 - 0.450915 | 0.000154 - 0.064382 | 0.0 - 0.218781 |
| 70s A2AP | 0.044737 - 0.925926 | 0.23129 - 0.450915 | 0.000448 - 0.009413 | 0.0 - 0.017295 |
| 70s T2TP | 0.007999 - 0.89899 | 0.181625 - 0.353692 | 0.000154 - 0.064382 | 0.0 - 0.218781 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attributes pt 2 | bipartite | BipartiteShortestPathto<60s,70s>Node | Bipartite Community | ProjectedCommunity | Playcount |
| 60s Bipartite | 0,1 | 0-5 | 1-306 | 1-427 | 0-507244941 |
| 60s A2AP | 0 | 1,3,5 | 1-306 | 1-427 | 0-507244941 |
| 60s T2TP | 1 | 0,2,4 | 1-306 | 1-9 | 0 |
| 70s Bipartite | 0,1 | 0-5 | 1-194 | 1-272 | 0-507244941 |
| 70s A2AP | 0 | 1,3,5 | 1-194 | 1-272 | 0-507244941 |
| 70s T2TP | 1 | 0,2,4 | 1-194 | 1-32 | 0 |









**IV. Analysis:** Compare the two networks using visualizations, tables, and other information. What are the similarities and differences?

The 60s network was larger overall with 11098 nodes to 8739 in the 70s network. In the 60s network, 87.8% of the nodes were artists, whereas in the 70s network that number is only 74.4%. The density of the two networks as a whole were similar; however, when looking at the projected networks, the artist network for the 70s is denser than that of the 60s, but the reverse is true when looking at the tag projected network.

It is possible to sort all the 60s tags to into only 9 communities, while the minimum number of communities is 32 for 70s tags. This implies the 60s network is more interconnected

The degree distribution of the nodes were similar to what we had expected. Except for artists which had close to degree 0, they followed a typical distribution. A lot of tags were only associated with a few artists, but few very common tags like ‘pop’ or ‘rock’ were highly connected, therefore the degree distribution for tags was as expected, with many tags having few connections and few tags having the majority of connections.

**V. Conclusion:** Summarize the findings.

In conclusion, we found that our hypothesis could not be proven correct. The 70’s network was not more spread out overall, and it was not split into more communities than the 60’s network. In fact, the evidence suggested opposite was the case at times.

**VI. Future Work:** Indicate additional questions raised by the research. What would you do if you had more time?

If we had more time, it would be interesting to see how these trends continue throughout the decades. Would 80s music be more segregated than 70s? If we had more time, we could make networks for all of the decade tags, filter more precisely, and then see if we could find more trends as time goes on.

It would be worth modifying the code so that we could delete duplicate nodes, or not allow them in the first place (aka nodes that have the same label). We might need to create a dictionary of labels:nodes and either delete the second node or not allow it to be created. We also could modify the code to allow for passing in the initial tag during initiation. Currently we need to manually comment-out the initial 60s nodes when crawling for 70s and vice versa, which is not sustainable. It would probably require modifying classes apart from my\_api, which is why we didn’t try to do it this time.

It would be interesting to see what happened to the graph if the number was changed to 20 or limited to 3, or if tags and artists had different numbers returned. Also, we noticed that when we added a query for a ‘playcount’ attribute for each artist, the time to crawl went up significantly. We did not do anything regarding edge weights for the original graph; however, we used the weighted\_projection method for artist-to-artist and tag-to-tag projections in step 3. The easy queries, as in the queries that were immediately returned by the pylast package, did not lend themselves easily to creating edge weights in any way. We also noticed that artists have additional information of some sort of weight with regards to the tags they are gathered from. For example, the first artist result from the folk tag would say Bob Dylan, 100, and second would say Simon and Garfunkel, 78. It may have been helpful to find a way to add that into the graph, and also calculate the correlation between playcount and all the other measures/clustering coefficients.