Analysing and Modeling Musician Networks

Leon Zhuang

March 2023

1 Abstract

Using a dataset of musician collaborations from Spotify, we find that the structure of genre and collaboration networks are dissimilar. We also find that there is homophily in that artists are more likely to collaborate with others of the same genre. Lastly, we attempt to model the collaboration network with a random graph with edge construction based on homophily and find it can explain the community structure of this network.

2 Introduction

Musicians interact with musicians, whether it is directly through a collaboration on a record or indirectly through remixing one another's song. Thus, it is natural to think about these interactions in terms of a network. Using networks constructed from a dataset of Spotify artist collaborations, we show that 1: the structure of genre and collaboration networks are dissimilar, and 2: artists tend to collaborate with artists of the same genre: in other words, homophily. Finally, we attempt to model the collaboration network as a synthetic network made with probabilistic edge construction with p based on our finding of homophily. We find that while our model does not completely capture the general structure of the collaboration network, it sufficiently simulates its community structure.

First, we review some of the existing work that revolves around our questions. In section 4, we review some concepts from graph and network theory used later in our empirical analysis. Section 5 and 6 discuss our methodology and empirical results and Section 7 discusses future work.

3 Related Work

Previous work has been done at the intersection of our proposed research direction and networks of musicians. Topirceanu et al. (2014) constructs and analyzes a network of musicians, to gain insights about communities, influential nodes, and similarities to other types of social networks. The authors compared MuSeNet to 5 other social networks, using six metrics that are sufficient to characterize the network: average degree (AD), average path length (L), average clustering coefficient (C), modularity (Mod), graph edge density (Dns) and graph diameter (Dmt) [1]. They also introduce a novel measurement called sociability, which measures whether one network is less or more sociable than another, which is a function of these six metrics [1]. They found that Musenet is less sociable compared to more traditional "friendship" social networks, such as Facebook and Twitter. The "real-world" explanation put forth by the authors is that musicians are less likely to form links (collaborate) with other musicians than people are to form friendships with other people.

Jacobson and Sandler (2008) also analyze a network of musicians, but construct the network using a different notion of connectivity. They construct the network using MySpace artist pages. Two artists are connected if either lists the other as a top friend on their personal page [2]. To obtain a sample network that is representative of the whole network, snowball sampling with $d = 6$ was used, resulting in 15,476 nodes [2]. The process involves choosing a root node and growing the network by adding nodes discovered by breadth-first search up to a distance of d from the root [2]. Subsequent analysis of this network revealed a small mean geodesic distance and diameter, which point to a small-world network structure [2]. The authors also aim to answer the question if this network exhibits homophily with respect to genre. In other words, are artists more likely to be friends with artists in the same genre? To answer this question, they first calculate the assortativity coefficient, which is a measure of homophily. They find a value of 0.216, which points to an existence of homophily [2]. The authors also find a negative correlation between geodesic distance and number of common genres.

Both of these papers examine networks of musicians, but with subtle differences. For instance, they take different approaches as to what the definition of a connection is. Topirceanu et al. connects two artists if they have belonged to the same band at some point in time, and the MySpace paper connects two artists if either lists the other as a top friend on their personal page. It turns out that the structure of these two networks is different. For instance, the clustering coefficient, C, for MuSeNet is 0.884, while for MySpace it is 0.219 [1][2], indicating a much higher presence of triangles in the former relative to the latter.

4 Basic Graph Metrics

These graph metrics are used in our analysis. Given an undirected graph G with a set of nodes V and the set of edges E these metrics are:

The average degree:

$$a = \sum_{u \in V} \frac{deg(u)}{|V|}$$

The average shortest path length:

$$L = \sum_{u,v \in V, u \neq v} \frac{dist_{min}(u,v)}{(|V|)(|V|-1)}$$

where $dist_{min}(u,v)$ is the shortest path from u to v,

The average clustering coefficient:

$$C = \frac{1}{|V|} \sum_{u \in V} \frac{2T(u)}{deg(u)(deg(u)-1)}$$

where $T(u)$ is the number of triads (three nodes all connected to each other) that node u belongs to,

The density:

$$Dns = \frac{2|E|}{|V|(|V|-1)}$$

The diameter:

$$Dmt = \max_{u,v \in V, u \neq v} dist_{max}(u,v)$$

where $dist_{max}(u,v)$ is the longest path from u to v.

The modularity:

$$Mod = \sum_{c=1}^{n} \left[\frac{L_c}{m} - \left(\frac{k_c}{2m} \right)^2 \right]$$

where n is the number of communities in G, m is the number of edges in G, L_c is the number of edges from community c to other communities, and k_c is the sum of the degrees of the nodes in community c.

These formulas are adapted from https://networkx.org/documentation/.

5 The Spotify Artist Collaboration Dataset

We use the dataset from https://www.kaggle.com/datasets/jfreyberg/spotify-artist-feature-collaboration-network, which contains information about 300,386 collaborations of 156,422 artists. It also contains information about which genres an artist works with.

6 Empirical Results

Note that all graphs in this paper are undirected graphs unless otherwise mentioned.

6.1 Comparing Different Methods of Construction

How does the structure of a network change as you change the definition of a connection? For networks of musicians, there are multiple musical connections possible between two artists. It is obvious that something will change if you reconstruct the network using a different notion of connectivity. For example, two artists that have remixed each other but have not co-produced a song would be connected in a “remix” network, but disconnected in a “co-production” network. Because this is obvious, it is more interesting to analyze how the overall structure of the network changes using holistic metrics similar to [1].

6.1.1 Network Construction

We used Python and the NetworkX library to conduct analysis on this dataset. Several of the metrics, such as shortest average path length, assume a connected network (because in a disconnected graph, the path length between nodes in two different connected components is undefined) so we conducted our analysis using the largest connected component. The size of the largest connected component was 52882.

However, from initial tests, we realized that metrics like average shortest path length and diameter took a non-trivial amount of time to compute. It makes sense since these metrics involve computing the longest/shortest paths between all pairs of nodes. As the sampled network is connected, there are $(V)(V-1)$ pairs of nodes. The time complexity would then be $O(V^2)$. After several hours, they had not finished yet, so we had to reduce the size of our sample. We used the snowball sampling discussed in [2] with $d = 4$, obtaining a network of 4742 nodes.

6.1.2 Direct comparison of basic graph metrics

The dataset contains enough information to construct two networks: a collaboration network and a genre network. In a collaboration network, two nodes are connected if they have co-produced a song, and in a genre network, two nodes are connected if they share at least one genre. After construction, the six common graph metrics were computed for each network:

Network Metrics						
Network	a	L	C	Dns	Dmt	Mod
Collaboration	14.452	3.429	0.183	0.003	6	0.424
Genre	217.560	2.824	0.783	0.049	10	0.375

There are some obvious differences here: namely, the average degree and average clustering coefficient. The explanation for an extremely high average degree in the genre network is that all artists who belong to a particular genre will be connected to each other. For instance, if there are 200 artists who have the label "techno", each techno artist will be connected to every other techno artist, meaning each techno artist has at least 200 edges. Since all techno artists are neighbors of each other, it means there are many triangles that exist in the network, explaining the C of 0.783.

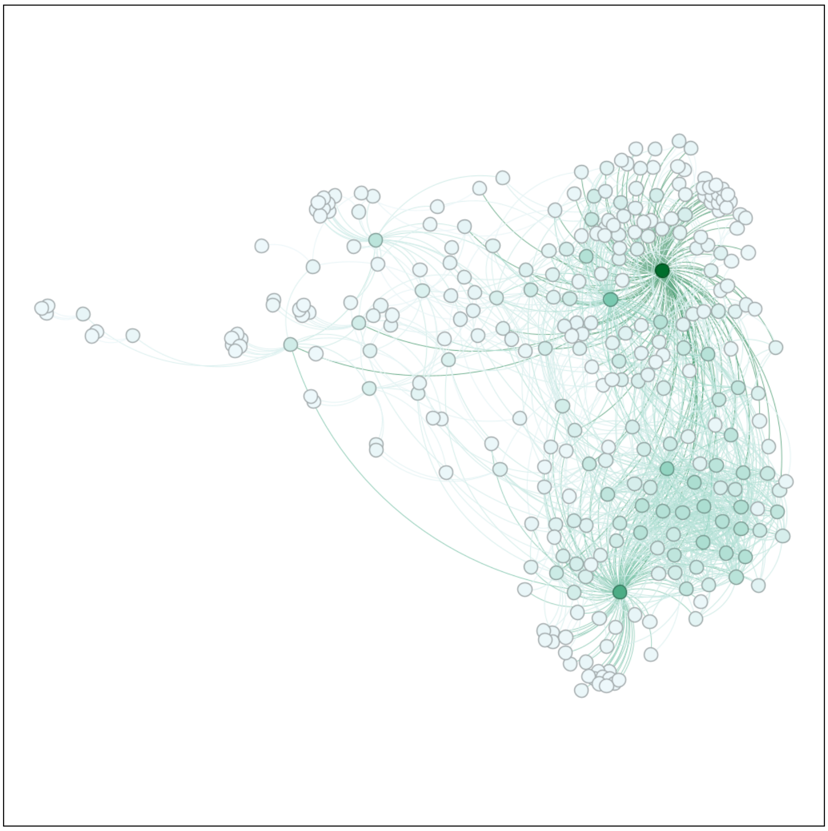

Figure 1: Collaboration network.

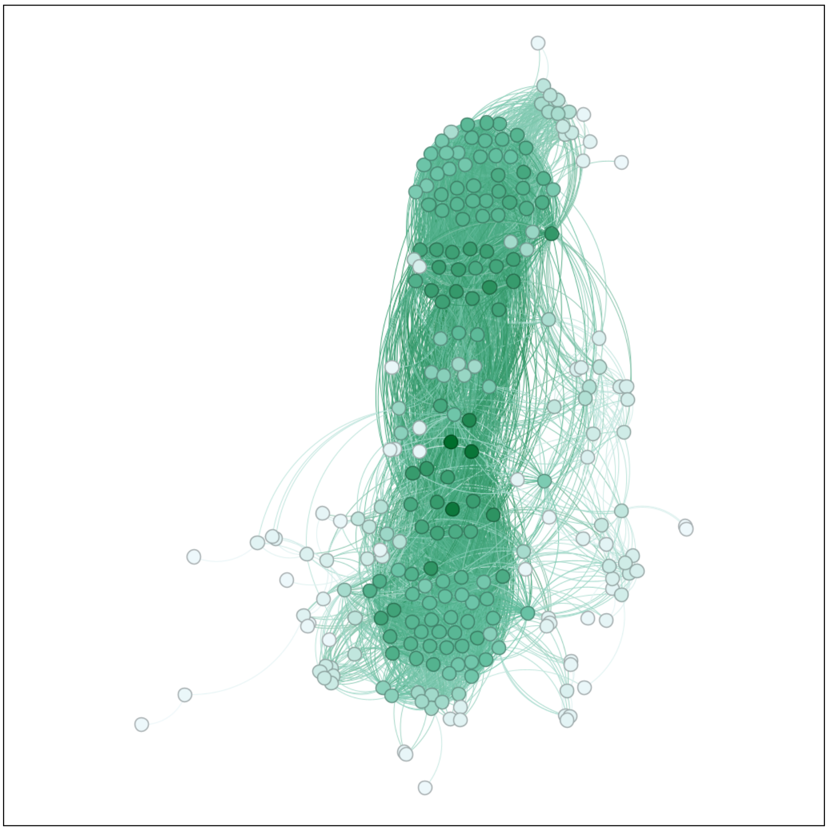

Figure 2: Genre network.

Using Gephi, we were able to visualize these two networks to visually support our explanation. Note that the depicted networks are only samples (around 300 nodes) of the actual networks obtained via snowball sampling. If the entire network was shown, individual nodes would be impossible to see, making it difficult to illustrate our point. The more dark green a node is colored as, the higher its degree is. There are significantly more dark green nodes in the genre network than in the collaboration network, which verifies the huge disparity in a between the two networks.

6.1.3 Variance of basic graph metrics

An alternative to direct comparison of the graph metrics between the different networks is to compute the variance of each graph metric, which is a more succinct way of determining whether or not the structure varies greatly from network to network. Granted, computing the variance is not as important to accomplish this when we are just comparing two networks, but makes it much easier when dealing with a large number of networks.

Specifically, the variance is defined as

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Variance of Network Metrics						
	a	L	C	Dns	Dmt	Mod
σ^2	20626.43	0.183	0.18	0.001	8	0.001

There is huge variance in a, which is not surprising, given what we have observed earlier. This is enough evidence to confidently say that these two networks differ greatly in structure.

6.1.4 Applications

One may wonder about the practical applications of a differently-defined network and measures of its structural variance. Graph-based music discovery and recommendation systems are one such real-world application. We envision a system that, given an artist that the user likes, allows the user to define how it recommends new artists. If the user wants to find artists that collaborated with their chosen artist, the system constructs the network of collaborations, and outputs the artists that have edges to the chosen artist. If the user decides they want to find artists that have remixed their artist, the system constructs the network of remixes. And so on. A user may also be interested in finding out how strongly musically connected their network is. To accomplish this, we could construct a network for each musical connection in the database, then compute the structural variance. If the variance is low, it means that on average, edges between two artists tend to persist across the different networks. This in turn means that artists on average are musically connected through most of the possible ways that they can be musically connected, indicating a ”strongly” musically connected network.

6.2 Homophily with Respect To Genre

Existing work has shown that musicians are more likely to form friendships with others of the same genre [2]. Here we ask a similar but different question: are musicians more likely to collaborate with other musicians of the same genre?

6.2.1 Assortativity Coefficient

We use the assortativity coefficient as defined in [2] to quantify homophily in this network. Let there be a graph G with a set of nodes V and a set of edges E. Additionally, let there be N unique labels $L_1, L_2, ..., L_n$ that a node can be associated with. Every node is associated with no more and no less than 1 label. Then an $N \times N$ matrix M can be defined such that $M[L_i][L_j]$ is the number of edges between nodes of label L_i and label L_j.

Then the normalized mixing matrix e is constructed such that

$$e = \frac{M}{|E|}$$

Finally, the assortativity coefficient is defined as:

$$a = \frac{Trace(e) - ||e^2||}{1 - ||e^2||}$$

a will be negative if nodes tend to connect with differently labeled nodes, 1 if nodes only connect with nodes with the same label, and 0 if there is no preference either way [2]. As there are 4838 unique genres in our network, M is a 4838 × 4838 matrix. Applying this calculation to our network, we obtain $a = 0.213$. Note that unlike the calculation of the graph metrics in section 5.1, this calculation does not use a network obtained from snowball sampling. The network used is obtained (let us call it G_1) after removing nodes with no genres associated with them, along with any edges associated with them. The result is a graph with 52842 nodes and 158169 edges. To see whether the value of a we have obtained indicates a significantly higher amount of homophily than natural, we compare it with a random graph G_2. This random graph is constructed by removing all edges from G_1. Then, an edge is created between a randomly selected pair of nodes. If an edge already exists, another pair is selected until one is found with no edge. This process is repeated until 158169 edges have been created. Calculating a for such a graph gives us a value of 5.84×10^{-05}. As the value of a for the real network is orders of magnitude larger, we can conclude that there does exist homophily in this network in terms of collaborations.

We can also verify our findings graphically. To accomplish this, we use a method used in previous work, which is to visualize the compatibility matrix of the dataset [3]. Let n_i be the number of outgoing edges from nodes of label L_i, Then let the compatibility matrix M' be an $N \times N$ matrix such that

$$M'[L_i][L_j] = \frac{M[L_i][L_j]}{n_i}$$

Essentially, $M'[L_i][L_j]$ is what percentage of collaborations involving artists of genre L_i are collaborations with artists of genre L_j. Thus, each entry of M' takes a value between 0 and 1. 1 indicates perfect association: artists only collaborate with artists of the same genre. If there is homophily, we should expect higher values along the diagonal of the matrix relative to other locations in the matrix. We can visualize this matrix as a heatmap:

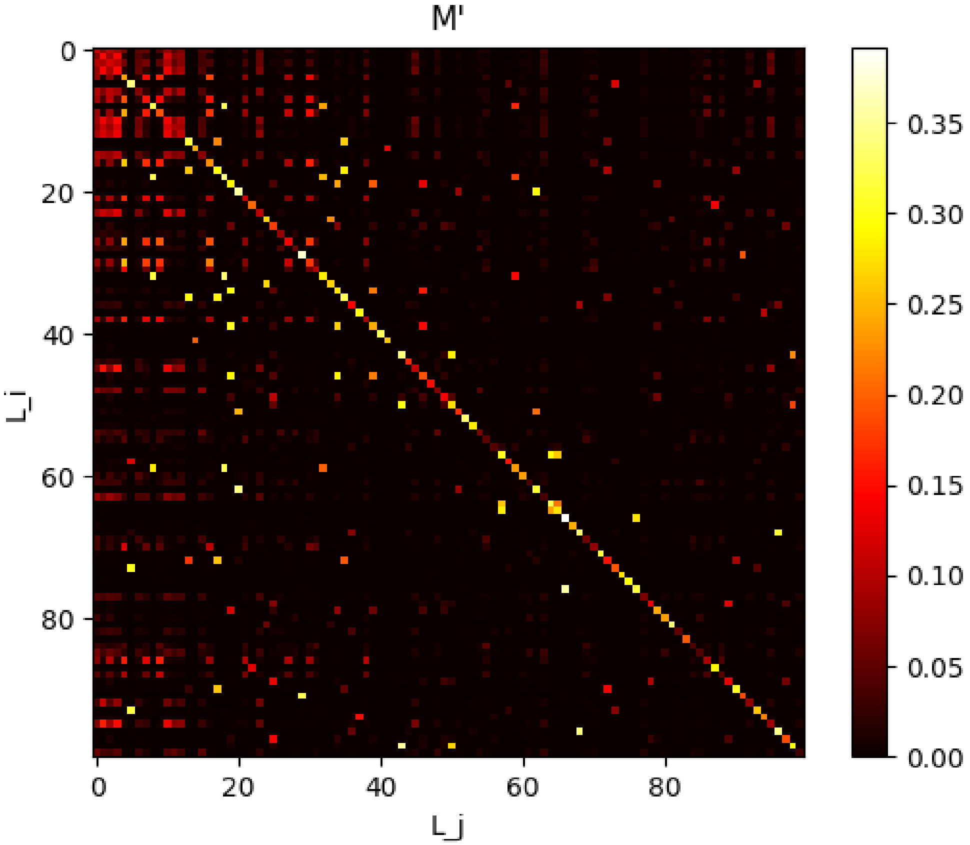

Figure 3: Heatmap visualization of M'

Because there are 4838 unique genres in total, visualizing the entire genre space would result in visually imperceptible elements, so we visualize the top 100 most common genres in the dataset. M' is ordered from most common to least common (top to bottom and left to right). As expected, there are high values on the diagonal. However, there are also many "hot spots" outside of the diagonal. Given $M[L_i][L_j]$ is a non-diagonal hotspot ($L_i \neq L_j$), there are several viable explanations. One: L_i and L_j are conceptually similar genres, which makes collaboration more likely (for example, RnB and hip-hop). Two: there are certain genres whose artists that tend to collaborate more outside of their own genre. A prime example is jazz artists [1]. In the case of jazz, we might see the collaborations spread out more on its corresponding row on the heatmap rather than concentrated on the diagonal. Or three: popular genres often collaborate with popular genres. This hypothesis is backed up by the top-left portion of the heatmap, since those are where the most popular genres are, and there are a high concentration of non-diagonal hotspots there.

6.3 Modeling the Collaboration Network

Can the collaboration network can be modeled using a random graph? To construct this graph, we begin by adding the first 5000 artists from the dataset as nodes. Then, we iterate through all unique pairs of nodes in the graph, and add edges with probability p. Our first strategy for choosing p incorporates the fact that there exists homophily with respect to genre (as we have found in the last section). For each pair, if they have at least one genre in common, an edge is added between them with $p = 0.03$. If they do not, an edge is added with $p = 0.003$. These low probabilities reflect the fact that even if two artists share common genres, the chance of collaboration is still very low. We compare the resulting network with the collaboration network constructed in Section 1.

Network Metrics						
Network	a	L	C	Dns	Dmt	Mod
Collaboration	14.452	3.429	0.183	0.003	6	0.424
Random 1	16.443	3.352	0.003	0.003	5	0.0

The values for a, L, Dns, and Dmt are very close, but there is a large disparity for C and Mod. This means the collaboration network can be divided into communities, and has more triangles, where as our model has very little triangles present and there is absolutely no community structure. Our intuition for near-zero values for Mod is that p is the same no matter what genres the two nodes are. This means all parts of the network have an equal chance of being connected, so there does not exist a part of the network that is more densely connected than any other part. This would result in very low modularity. As for C, triangles are simply very unlikely to exist in our model. The highest likelihood for a triangle to exist is when all 3 nodes share a common genre. Even then, the probability for a triangle is $0.03^3 = 2.7 \times 10^{-5}$.

All of this points to needing a more dynamic and generally higher p. Our second model accomplishes this by adding more nuance to the homophily aspect, by taking into account the fact that genres do not have equal probabilities of collaborating with each other. We determine p as follows:

Let $L_u = \{L_{1_u}, L_{2_u}, ..., L_{n_u}\}$ and $L_v = \{L_{1_v}, L_{2_v}, ..., L_{m_v}\}$ be the sets of genre labels associated with nodes u and v, respectively. Then p is defined as

$$p = \max_{L_{m_v} \in L_m, L_{n_u} \in L_n} M'[L_{m_v}][L_{n_u}]$$

We can view a collaboration between node u and node v as a collaboration between each of their associated genres. For example, if node u is associated with Pop and Rap, and node v is associated with Hip-Hop and Pop, there are 3 genre collaborations: Pop and Hip-Hop, Pop and Pop, and Rap and Hip-Hop. These correspond to three different entries in the compatibility matrix, and we simply take the maximum value for p.

Network Metrics						
Network	a	L	C	Dns	Dmt	Mod
Collaboration	14.452	3.429	0.183	0.003	6	0.424
Random 2	25.829	3.463	0.135	0.005	8	0.631

The structures of the two networks are much closer this time. The biggest improvements are the metrics that we have aimed to change, C and Mod. It means that our homophily-based model can adequately model the community structure of the real network. However, it is still not a good enough approximation of the general structure of the network. a seems to be a bit too high, which means connections are formed more often in our model than the real network. This means that p is on average too high. We believe that this is due to what the compatibility matrix measures. To recap, $M'[L_i][L_j]$ is what percentage of collaborations involving artists of genre L_i are collaborations with artists of genre L_j. When deciding whether two nodes should be connected, we want the likelihood of L_i to collaborate with L_j independent of total amount of collaborations, so these two probabilities are not the same. Future work will involve finding a more accurate method to measure the likelihood of one genre to collaborate with another.

7 Conclusion and Future Work

Analysis of the Spotify artist dataset reveals several key insights. First, constructing networks using different definitions of an edge results in great variance in their resulting structures, case in point being the collaboration and genre networks. Second, genres have different likelihoods of collaborating with each other, and in particular, there is homophily in that genres are more likely to collaborate within themselves than with other genres. Finally, a random graph model utilizing homophily is able to emulate the community structure of the collaboration network, allowing us to conclude that homophily is an good explanation for the community structure found in the collaboration network.

7.0.1 Future Work

In section 5, we explored the question of how the structure of musician networks changed as we redefined what it meant to have an edge between two artists. We measured this amount of change by computing the variance of basic graph properties on the set of differently-defined networks. Robust statistics need a sufficient sample size, and there are two places in our methodology where the sample size could be increased.

Computing the variance for each graph metric used a sample size of 2, as the chosen dataset contained information about only 2 types of musical connections (thus limiting us to the construction of only 2 different networks). This sample size is not ideal.

Although this dataset was not ideal, we gauged it to be the best option at the time due to several factors. First, we were unable to find another pre-assembled dataset that contained more than 2 types of connections. This meant any better datasets would have to be manually assembled. If we decide to go this route, the best source for this data would have been WhoSampled.com, a database dedicated to recording the different musical connections between artists. However, we were unable to gain access to their API. To get around this, we could use web scraping, but it would be much slower and extremely inconvenient and we would have to be careful not to send too many requests to avoid getting IP banned. In short, we realized that manually assembling the data would be a major undertaking, and it would be unfeasible due to time constraints.

For future work, we aim to assemble a dataset that contains at least 5 types of musical connections. Other than collaborating or producing the same genre, artists could also belong to the same band, sample each other's songs, or appear on the same release. A sample size of 5 allows us to do a more robust variance calculation.

Perhaps it is more convenient to be able to tell at a glance how structurally similar a set of graphs are. One possible solution is to create a single metric that measures this similarity. There is precedence for this sort of approach; the ”sociability” measure proposed by [1] is a function of six graph properties. However, the sociability measure is not a good fit for our problem, as it only takes two graphs as input and weighs certain metrics differently based on whether they contribute positively or negatively to sociability. We need a metric to compare an arbitrary number of networks. Here we introduce a ”rough draft” of such a metric called the structural variance metric.

Let there be a set of graphs $S = \{G_1, G_2, ..., G_n\}$, and a set of graph metrics $M = \{M_1, M_2, ..., M_n\}$ such that $|S| > 1$ and $|M| > 0$.

The structural variance of this set, SV is defined as

$$SV = \sum_{m \in M} Var\{m_g \mid g \in S\}$$

where m_g is a certain metric of a graph g.

In other words, for each metric, we compute it for all graphs in the set, obtaining a set of values, and find the variance of this set. Then, all the variances are summed up.

For example, taking the values we have found in section 5, we obtain $SV = 20634.80$. Interpretation of this metric is simple: the farther away from 0 the value is, the more structural dissimilarity in the set of graphs. The closer the value is to 0, the less structural dissimilarity in the set of graphs. However, if the value is 0, it cannot be said that all the graphs are structurally identical i.e. the adjacency matrices of all the graphs are the same. Note that M can contain an arbitrary number of metrics. This means there could be a metric that is not used in the calculation of SV which could vary among the set, indicating that the set of graphs are not identical. This means SV measures similarity only with respect to the metrics in M.

There are various places for improvement with current iteration of this metric, which is why it is introduced in this section. Specifically, there is no normalization on the set of metrics before calculating the variance. Certain metrics operate at different scales: for an empirical proof, we can take a look at the network metrics computed in section 5. a gets into the scale of the hundreds, while Dns, C, and Mod are between 0 and 1. The difference in scale means it is not possible to have a meaningful comparison between different values of SV without normalization. A direction worth looking into would be the normalization method discussed in [4].

8 References

1. A. Topirceanu, G. Barina and M. Udrescu. ”Musenet: Collaboration in the music artists industry”. IEEE, (2014)

2. K. Jacobson and M. Sandler. ”Musically meaningful or just noise? An analysis of on-line artist networks”. Proc. CMMR, (2008), pp. 306-314

3. Lim, D., Hohne, F., Li, X., Huang, S. L., Gupta, V., Bhalerao, O., Lim, S. N. ”Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods”. Advances in Neural Information Processing Systems, (2021), 34, 20887-20902

4. A. Topirceanu, M. Udrescu and M. Vladutiu. ”Network Fidelity: A Metric to Quantify the Similarity and Realism of Complex Networks”. 2013 International Conference on Cloud and Green Computing, Karlsruhe, Germany, (2013), pp. 289-296, doi: 10.1109/CGC.2013.53.