Extracting and analyzing the artist graph on Spotify

Tara Morovatdar
University of Koblenz-Landau

The music industry is a billion dollar industry and a source of interaction between individuals. It is not just interaction between artists and listeners but also among artists. In this paper, we are going to analyze artist-artist relations and then we will answer a few questions regarding interesting patterns found in the data. In this paper, we extract the network of artists in Spotify in which the nodes are the artists and the edges show the relation between artists. 1- We are going to explore the relationship between the network structure (degree centrality, closeness centrality, betweenness centrality, and page rank) and being successful (One way of measuring how successful is someone is to measure the popularity of the artist).
2- We are going to detect communities in the network and find how and based on what artists belongs to these communities, using Louvain algorithm.

Additional Key Words and Phrases: Network structure,artist graph,artists on Spotify, Centrality Measures, Louvain, Community detection

1. INTRODUCTION

Spotify is a music, podcast, and video streaming service that started in 2008. It is developed by startup Spotify AB in Stockholm, Sweden[Wikipedia 2017c]. As it has Over 140 million (as of June 2017) active users and about 2 million artists[Spotify 2017] which makes it a big database that can be interesting for study.
Spotify allows clients to retrieve some information about artists and other artists related to them. The relation between artists can be projected into a graph where links determine which artists are related to each other. Usually, this information is extracted from the user's habits of listening to music. It can also facilitate finding new artists. "Spotify's related artists are determined by algorithms which look at what people listen to alongside your music. So if your music were to be put in a playlist alongside artist X and artist Y then artists X and Y are more likely to be shown as related to you or played on the radio."[Spotify 2015].
By generating the artists collaboration network, we are going to indicate their status in the music industry. Discovering how successful an artist is, as a part of this big industry has a high value. We are going to explore the relationship between the network structure and being successful. One way of measuring success is to measure the popularity of the artist. Spotify meta-data about an artist includes a measure called popularity, the popularity is a value between 0 and 100, with 100 being the most popular. The artist's popularity is calculated from the popularity of all the artist's tracks [Spotify 2016]. In this work, we first investigate the relationship between network structure and popularity. This will help us determine the nature of these two features. After that, we explore the underlying structure of the network by applying the Louvain clustering algorithm to uncover some communities in our data.
The organization of the paper is as follows. Section 2 introduces some of the related works, they include researches associated with both systems which try to solve each problem. Section 3 explains the suggested method and goes into the working details of each system. Section 4 concludes the paper.

2. RELATED WORKS

Music industry nowadays is a very large competitive market. It is very important and crucial to predicting whether an artist will be commercially successful in the future or not. Different machine learning techniques and measurements are used to predict the successful artist. One metric to measure how successful is an artist can be the number of followers in social media or the popularity. There have been several studies done to indicate the relation between network structure and socio-metrics status. In [Gest et al. 2001] they have shown that children with more social network centrality has a higher chance of entering an existing group and are more liked. John Cardente has used degree centrality, closeness centrality and betweenness centrality to identify the key innovators in an innovative culture [Cardente 2012]. Suzanne Stathatos and Zachary Yellin-Flaherty used degree centrality, betweenness centrality and Eigen centrality with a combination of a set of other features to train different machine learning algorithms and compare them to evaluate which method can predict the success of an artist better[Stathatos and Yellin-Flaherty 2014]. [Ren et al. 2014] assumed that a node with high centrality value usually occupies a crucial position in the social network, therefore, it has high potential to influence more other users." therefore they have considered five centrality measures to identify the influential users. There are several community detection algorithms :divisive algorithms which have top-down approach like GirvanNewman algorithm which removes edges from original network and remaining connected components are communities[Wikipedia 2017b], agglomerative algorithms which have bottom-up approach use a measurement to find the similarity between vertices and then groups iteratively the vertices into communities[Pons and Latapy 2006] and optimization methods which tries to maximize an objective function. [Blondel et al. 2008] The measure that is used to compare these algorithms is called modularity of the partition. The modularity of a partition has a value between - 1 and 1 that measures the density of links inside communities as compared to links between communities.[Newman 2006].
Detecting communities in networks has become a fundamental problem in network science. Many algorithms have been proposed. In this paper, we have used the Louvain algorithm to detect communities. Louvain is a method to extract the community structure of large networks. It is a simple heuristic method that is very fast in comparison to other known community detection methods and moreover, the quality of the communities detected base on modularity is very good.[Blondel et al. 2008] The algorithm is divided into two phases that are repeated iteratively. First, it starts with a network of N nodes and assigns a different community to each node of the network. Then, for each node, it calculates the gain of modularity if we remove the node from its community and by placing it in the community of its neighbor. The node is then placed in the community for which the gain is maximized. If no positive gain is possible, the node stays in its original community. This process is applied repeatedly for all nodes until no further improvement can be achieved and the first phase is then complete.In the second phase, it builds a new network which the

nodes are the communities from the first phase and the wights of the links are the sum of the weights of the links between tow communities and the links between nodes from same community are considered as self-loops. When the new network is generated it applies the first phase again and it iterates until there are no more changes and a maximum of modularity is gained.[Blondel et al. 2008]

3. METHODS

Here, in the first section, we try to determine the relation between network structure and being successful.

3.1 Constructing the Network

In this paper, we first aim to investigate the topological structure of the artist graphs on Spotify. We do so by mapping the artists in a directed graph, in which links represent a connection between two related artists according to Spotify.
We Only consider the first 5 related artist and all the links in the graph have weight 1. We start with an initial list containing 70 artists, especially those who are famous, also we tried to choose from different genres. Then by utilizing a sampling algorithm, we construct the network. Various sampling methods have pros and cons. There are two goals of sampling, one is to scale down the graph and the other is to sample in a way that it is similar to the original one back in time. Since investigating the whole graph of artists on Spotify is very expensive in terms of time and capacity, our goal is to scale down a large graph and get a similar sample that is able to preserve as many properties possible as the original network. We choose the Iranian market as our sample network since we believe we are more familiar with both the culture and the market.

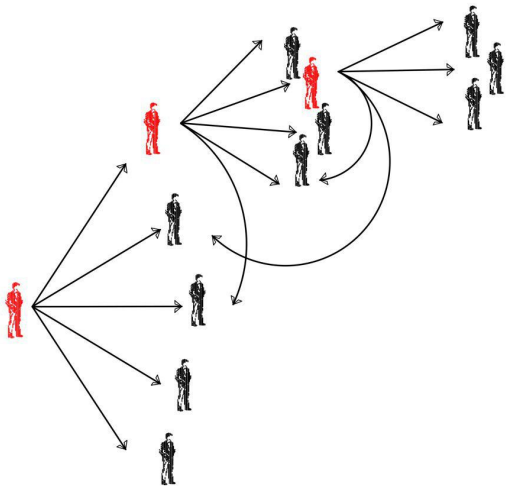

Fig. 1: How artist are related in Spotify. In each step the algorithm may find some visited nodes, if one of them is selected as the next step it will not be added to the network again.

There are different classes of sampling algorithms: 1- Node base 2- edge base 3- Crawling base techniques[Frank et al. 2012].There are a variety of techniques in each group as well. The same case is true for evaluation methods, which are going to be discussed briefly. We focus more on the approaches more qualified for the scaling down goal.
1- Node base algorithms:
-Is to uniformly and randomly choose a subset of nodes as our sample, this is called Random Node(RN).In [Stumpf et al. 2005] it has been shown that this method does not preserve some attributes of the original graph like degree distribution.
- Another approach is not using uniform probabilities but to set them based on the page rank weight, this is called Random Pagerank Node(RPN)
- Another technique is to specify the probability for each sample proportional to its degree which is called Random Degree Node(RDN) [Adamic et al. 2001].
2- Edge base algorithms:
- The simplest idea is called Random Edge(RE). This method selects edges uniformly which may end in a sparse graph. So the attributes like the community structure and the diameter will not match [Leskovec and Faloutsos 2006].
- The Second algorithm is called Random Node Edge(RNE) where you pick nodes randomly and then edges connected to the selected node are again chosen arbitrarily.
- The Third method is called Hybrid which is a special combination of RNE with a probability of p and RE with probability 1-p [Krishnamurthy et al. 2005].
3- Exploration base: In this type of algorithms, we are supposed to select a set of nodes and explore the nodes adjacent to them.
-Snowball or chain referral sampling is a simple method that has been widely used in different research areas. The procedure of snowball sampling is defined as below: a random sample is drawn from a finite population then in each step every sample will introduce K samples and from that K we take the samples(the order can be in BFS or DFS) that have not seen before and iterate until a number of defined steps has been reached or a predefined number of samples has been achieved[Goodman 1961].
- Random Node Neighbor(RNN) in which we choose a node and all its outgoing links uniformly at random [Leskovec and Faloutsos 2006].
- Forest Fire is another algorithm described below: in the basic model in each step at node i, we choose a node j and then generates a random number k and choose k links among in-links and out-links of j and it iterates [Leskovec et al. 2005].This algorithm is more proper to go back in time goal and not scaling down [Leskovec and Faloutsos 2006].
- Random Walk in which we uniformly choose a node at the start of the random walk and with probability p we go back to the start point. There is the problem of getting stuck in a small connected component and not visiting enough nodes. For this problem there is a solution that we set a threshold k and if in K steps we don't generate a new node we choose a different node as a starting point. In this paper, as we will discuss we applied this method for sampling from Spotify but to prevent sticking we have an initial list of seeds that we pick the starting node from those at uniformly random.
Random jump is another algorithm which is similar to random walk but with probability p we jump to any another node.
Researches based on practical conclusions have revealed that random walk is the best algorithm for scaling down purposes, since it is biased toward high degree nodes it gives connected sampled graphs which have similar properties for a set of representative properties such as degree distribution, distribution of clustering coefficient, community structure, etc[Leskovec and Faloutsos 2006]. In this paper, the process starts with an artist from the initial list. The initial list of artists includes 70 artists on Spotify from different genres

and ages. Constructing the network in each step, for the artist i, we find five related artists to the artist i then with the probability p=0.8 we select one of these five as the next candidate or with probability p=0.2 we go back to our initial seeds to choose our next node. If one of the related artists is chosen; if it is a new artist it is added to the network or if there is no link starting from artist i to the new artist then a directed link will be drawn otherwise the random walker go to the beginning of the process and iterate over all. We repeat this processes till 200 nodes(artists)are in the network. This number is chosen since it seems quite enough and includes most of the artists in the market and if we continue further the process will add non-Persian artists.

In the end, the network consists of 200 nodes and 839 edges. Last but not least it has to be considered that because of the small world property of the network as the network grows the probability of finding a new artist gets smaller in each iteration. As you can see in Fig 2 the amount of time needed to add new nodes to the network grows exponentially.

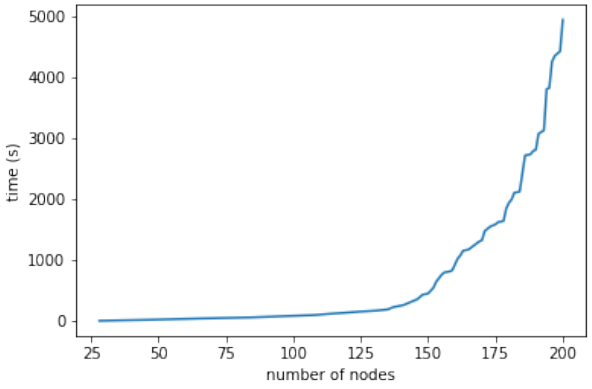

Fig. 2: The amount of time it takes to construct the network.

3.2 Network structure analysis

Employing our random walk algorithm, in each step, we generate a request and send it to Spotify web API which includes the ID of the artist; in return, it will give us an artist object which has information about the artist (name, image, genre, popularity, number of followers,).The number of followers and popularity are the two values that we have used in our analyses. The popularity is one of the values that we are interested in. This value will be between 0 and 100, with 100 being the most popular. The artist's popularity is calculated from the popularity of all the artist's tracks.[Spotify 2016] . Considering these two values as a sign of being the successful artist we aim to compare them with a set of parameters in the network to identify the relationship between those parameters and being successful.The set of parameters that we have been investigating are as below:

1-Indegree Centrality: The number of headends adjacent to a node is called the indegree of the node. In the network since outdegree of a node can be from 0 to 5 and it does not represent anything for the simplicity, just the indegree centrality has been calculated for each node. It is used for finding very connected individuals, popular individuals, individuals who are likely to hold most information or individuals who can quickly connect with the wider network.[cambridge intelligence 2018]

2-Betweenness Centrality: a measure of centrality in a graph based on shortest paths. The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex. Nodes with high betweenness may have considerable influence within a network because more information will pass through that node. They are also the ones whose removal from the network will have the potential to disconnect graphs.[Wikipedia 2017a]

3- Closeness Centrality: Closeness centrality of a node is the reciprocal of the sum of the shortest path distances from the node to all n-1 other nodes. Since the sum of distances depends on the number of nodes in the graph, closeness is normalized by the sum of minimum possible distances n-1[Developers 2015]. It is usually used for finding the individuals who are best placed to influence the entire network most quickly.

4- Page Rank: Page Rank computes a ranking of the nodes in the graph G based on the structure of the incoming links. It is a variant of Eigenvector Centrality. Eigenvector centrality computes the centrality for a node based on the centrality of its neighbors. Since Page rank also takes link direction and weight into account it suits our network better. It reveals the nodes whose influence extends beyond their direct connections into the wider network[cambridge intelligence 2018].

Fig 3 shows how indegree centrality, betweenness centrality, closeness centrality, and page rank are correlated with popularity. In all four plots, it is clear that as the centrality grows the popularity grows also since there is no point in the lower part of the figures.

To study the correlation between network indicators and successfulness(popularity and number of followers) three different algorithms have been applied.

1-Pearson correlation coefficient: is a measure of the linear correlation between two variables. It is calculated by, the covariance of the two variables divided by the product of their standard deviations. It has a value between +1 and 1, where 1 is the total positive linear correlation, 0 is no linear correlation, and 1 is total negative linear correlation[Wikipedia 2018].

2- Spearman's rank correlation coefficient The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships. The Spearman correlation between two variables will be high when observations have a similar rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully opposed for a correlation of 1) rank between the two variables.

3- Kendall rank correlation coefficient: is the similarity of the orderings of the data when ranked by each of the quantities. It is calculated by the number of concordant pairs minus the number of discordant pairs divided by the total number of pair combinations. Its value is between -1 and 1. It will be high(i.e. 1) when the observations have a similar rank.

4- Average overlap: Is an idea which is based on the set intersection. sets are just bag of items so there is no rank but the Average overlap is a way of using set intersection for ranked lists. In general, the idea is to determine the fraction of content overlapping at different depths. In this research, we are going to compare for the top 20 artists(depth=20) in each list, since the top 20 successful artists have most of the market shares.

Each of these metrics has pros and cons, Pearson measures the strength of the correlation between two variable, on the other hand, in Spearman, the values can be ordinal and not just intervals and

[0]It is not so clear how Spotify assign this score but it is likely that an artists popularity score depends on how many people listen to their songs.

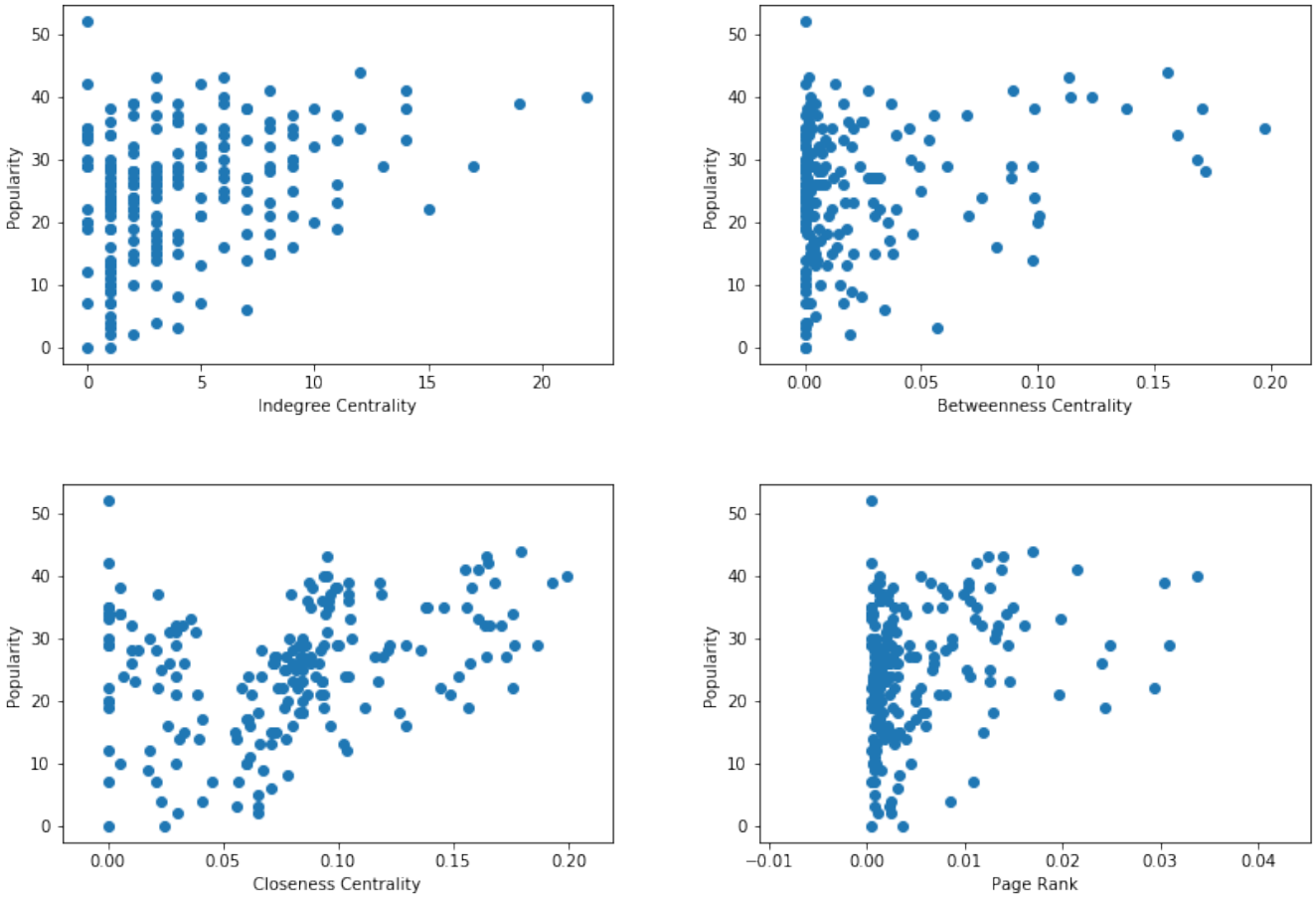

Fig. 3: Network metrics comparison with artist's popularity

Network Metrics	Pearson Correlation		Spearman's rank Correlation		Kendall tau Correlation		Average Overlap	
	Popularity	Followers	Popularity	Followers	Popularity	Followers	Popularity	Followers
Indegree Centrality	0.31	0.28	0.28	0.29	0.21	0.21	0.15	0.27
Betweenness Centrality	0.24	0.32	0.17	0.22	0.12	0.15	0.22	0.25
Closeness Centrality	0.32	0.23	0.35	0.25	0.25	0.18	0.23	0.25
Page rank (alpha=0.9)	0.28	0.25	0.24	0.20	0.17	0.15	0.15	0.12

Table I. : Network metrics correlation to the artist's popularity and number of followers

also it is not required for the relationship to be linear.[Hauke and Kossowski 2011]. Although Kendall tau is a measurement of rank correlation, it calculates the probability of a pair being in the same order and doesn't give priority in cases where there are concordant matches in the lists. Average Overlap(AO) gives more weight to the top of the list but still, there is a problem with the list being short, it gives out smaller numbers[Ekstrøm et al. 2015].The results are shown in table 1. Although the numbers are small in general, all of them show a positive correlation. Indegree centrality and closeness centrality have a higher correlation in general and Kendall tau has produced smaller values.

3.3 Artist Network Clustering

One of the most important features of graphs representing real network is community structure or clustering. The organization of nodes appears in clusters, in a way that many internal links are in each cluster and comparatively few links connect nodes of different clusters. Such clusters, or communities, can be considered as fairly independent compartments of a graph. Many algorithms have been proposed to find these communities[Lancichinetti and Fortunato 2009]. In this paper, we have used the Louvain algorithm to detect communities. Louvain is a method to extract the community structure of large networks. It is a simple heuristic method that is very fast in comparison to other known community detection meth-

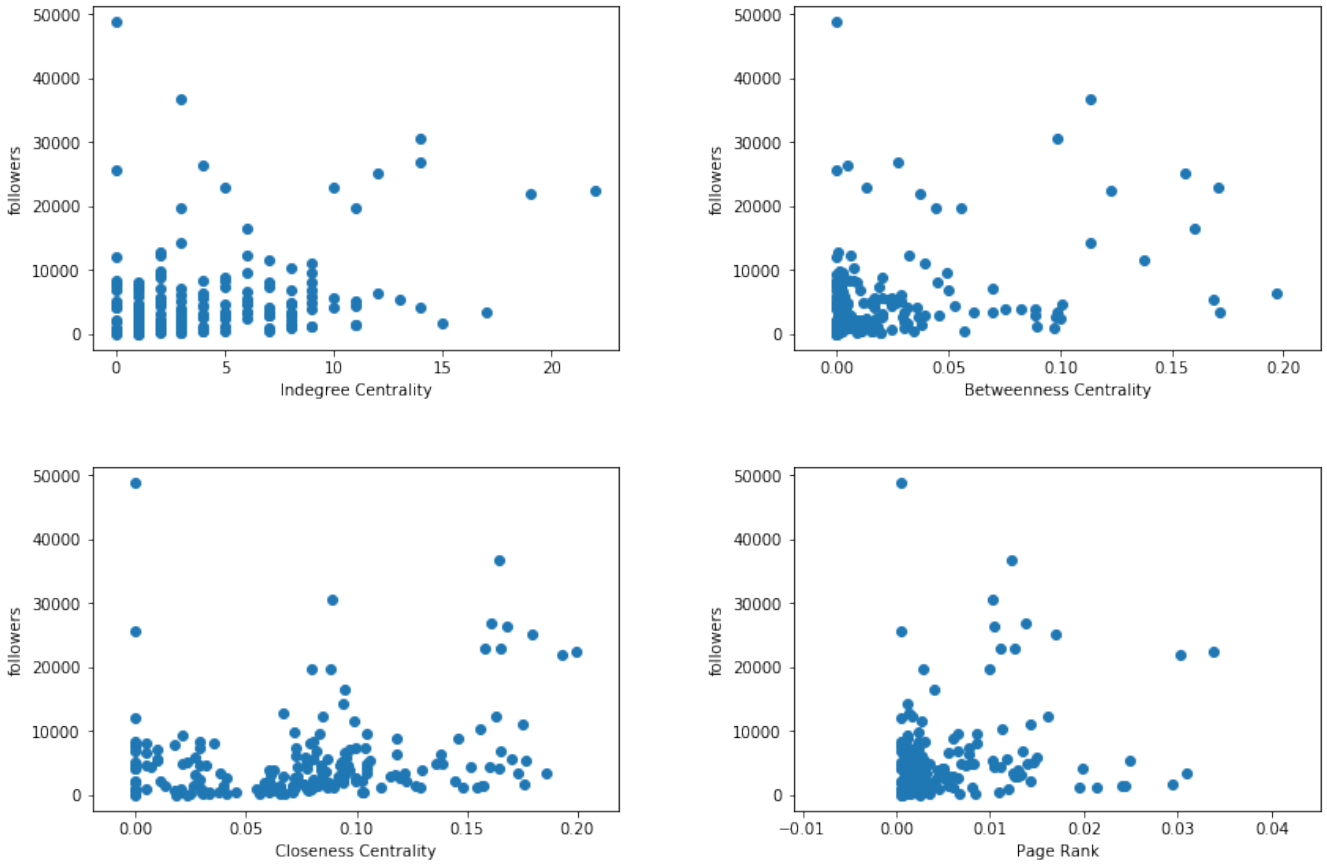

Fig. 4: Network metrics comparison with artist's number of followers

ods and moreover, the quality of the communities detected base on modularity is very good[Blondel et al. 2008].
Applying the Louvain algorithm on the whole network 9 communities were found with the modularity of 0.736. In fig 5 you can see the communities and the artists name. Since the general graph is not really representative, nodes with the degree less than 7 have been filtered. Finally, 60% of the network which includes 121 nodes and 551 edges are shown in fig 5. Also, it has to be noticed that nodes are rescaled proportionally to their degree so the nodes with a higher degree have shown with bigger circles. Based on our knowledge on Persian music market it can be seen that the communities are formed mostly based on the genres. Artists with the songs from the same genre are apparently clustered in the same group or they are in 2 different but connected compartments. Cluster 1 and 2 includes singers from classical Iranian music and as it can be seen cluster 1 is only connected to cluster 2. Cluster 3 includes singers who sing rap music. Cluster 4,5,6,7 and 8 can be categorized as Persian pop music but the interesting part is compartment 5 includes the young artists who perform inside Iran. Cluster 6 is the biggest component which includes a variety of pop singers, especially the best known and popular ones. Cluster 8 mostly belongs to the singers who performed some years ago. It is noticeable that group 9 (the yellow one) consists of singers of different genres like Rock and Rap. Since there are some artists in the group who are some kind of innovator in the Iranian market (i.e. Mohsen Namjoo) and as it can be seen this component is connected to different groups.

4. CONCLUSION

We have investigated the relationship between some network indicators and being a successful artist. As it is shown in fig 3 the correlation between popularity and different network metrics are positive; as the popularity is bigger the artist in more central. It might be interesting in all of the metrics that there are some artists who have high popularity and they are very successful but still they are not central. One of the plausible explanations can be that regarding the way that Spotify defines related artist; 2 artists are related if they appear often in the same playlist together, but users tend to create albums of their favorite artists alone. Basically, they are so popular that usually appear alone in an album and thus their relationship to the other artists is very meaningful.
In table 1 it has to be noticed that indegree centrality and closeness centrality have a higher correlation in general. Artists with higher indegree centrality have been appeared more often with other artists in a list the reason might be that they have collaborative songs or in general, their songs cover a range of different users taste.
During recent years, with the rapid growth of digital music collection, the need for an effective method to organize these data has

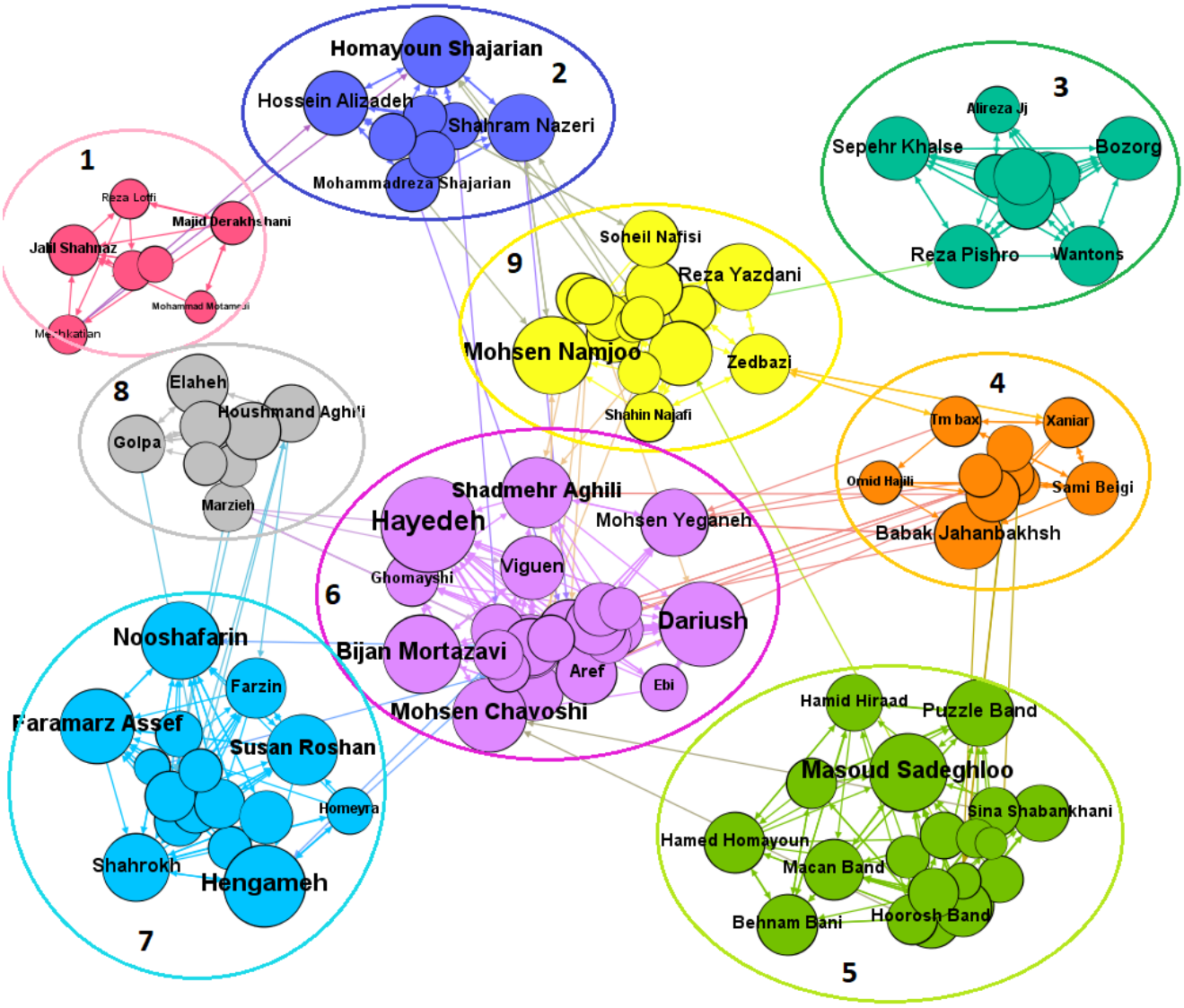

Fig. 5: Communities detected in the Network with the Louvain algorithm

been more noticeable. Music Information Retrieval (MIR) is an interdisciplinary research area which involves different approaches to organize the music collections such as music recommender systems, automatic music transcription, and automatic categorization[Wang 2016]. Through the community detection, we found some interesting patterns that are useful in the genre classification of artists.

Last but not least it has to be considered that since Spotify is censored inside Iran, not many people use it and if so the users are mostly from the young generation. This might bias the result in favor of the young generation taste. As a special group of people(Iranian who leave abroad, young people) have access to the platform this can have an impact on popularity and number of followers of the artists, thus identification of successful artists base on just only one platform may not be very precise.

REFERENCES

Lada A Adamic, Rajan M Lukose, Amit R Puniyani, and Bernardo A Huberman. 2001. Search in power-law networks. *Physical review E* 64, 4 (2001), 046135.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

cambridge intelligence. 2018. Social Network Analysis. (2018). Retrieved September 4, 2018 from https://cambridge-intelligence.com/keylines-faqs-social-network-analysis/

John Cardente. 2012. Using Centrality Measures to Identify Key Members of an Innovation Collaboration Network. (2012). http://snap.stanford.edu/class/cs224w-2012/projects/cs224w-043-final.pdf

NetworkX Developers. 2015. closeness Centrality. (2015). Retrieved September 4, 2018 from https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.centrality.closeness_centrality.html

Claus Thorn Ekstrøm, Thomas Alexander Gerds, Andreas Kryger Jensen, and Kasper Brink-Jensen. 2015. Sequential rank agreement methods for comparison of ranked lists. *arXiv preprint arXiv:1508.06803* (2015).

Dan Frank, Zhiheng Huang, and Alvin Chyan. 2012. Sampling A Large Network: How Small Can My Sample Be?

Scott D Gest, Sandra A Graham-Bermann, and Willard W Hartup. 2001. Peer experience: Common and unique features of number of friendships, social network centrality, and sociometric status. *Social development* 10, 1 (2001), 23–40.

Leo A Goodman. 1961. Snowball sampling. *The annals of mathematical statistics* (1961), 148–170.

Jan Hauke and Tomasz Kossowski. 2011. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae* 30, 2 (2011), 87.

Vaishnavi Krishnamurthy, Michalis Faloutsos, Marek Chrobak, Li Lao, J-H Cui, and Allon G Percus. 2005. Reducing large internet topologies for faster simulations. In *International Conference on Research in Networking*. Springer, 328–341.

Andrea Lancichinetti and Santo Fortunato. 2009. Community detection algorithms: a comparative analysis. *Physical review E* 80, 5 (2009), 056117.

Jure Leskovec and Christos Faloutsos. 2006. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 631–636.

Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2005. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 177–187.

Mark EJ Newman. 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences* 103, 23 (2006), 8577–8582.

Pascal Pons and Matthieu Latapy. 2006. Computing communities in large networks using random walks. *J. Graph Algorithms Appl.* 10, 2 (2006), 191–218.

Jing Ren, Zhiyong Cheng, Jialie Shen, and Feida Zhu. 2014. Influences of influential Users: An empirical study of music social network. In *Proceedings of International Conference on Internet Multimedia Computing and Service*. ACM, 411.

Spotify. 2015. Spotify Community. October 2015. (2015). Retrieved November 13, 2017 from https://community.spotify.com/t5/Content-Questions/quot-Related-artists-quot-how-do-Spotify-decide-what-artists-are/td-p/1185597

Spotify. 2016. Spotify API. (2016). Retrieved November 14, 2017 from https://developer.spotify.com/web-api/get-artist/

Spotify. 2017. Spotify Press. (2017). Retrieved November 8, 2017 from https://press.spotify.com/us/about/

Suzanne Stathatos and Zachary Yellin-Flaherty. 2014. (2014).

Michael PH Stumpf, Carsten Wiuf, and Robert M May. 2005. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences of the United States of America* 102, 12 (2005), 4221–4224.

Shu Wang. 2016. Musical Genre Categorization Using Support Vector Machines. (2016). http://homepages.cae.wisc.edu/~ece539/fall13/project/WangShu_rpt.pdf

Wikipedia. 2017a. Betweenness centrality. (2017). Retrieved September 4, 2018 from https://en.wikipedia.org/wiki/Betweenness_centrality

Wikipedia. 2017b. GirvanNewman algorithm. (2017). Retrieved march 9, 2018 from https://en.wikipedia.org/wiki/Girva-Newman_algorithm

Wikipedia. 2017c. Spotify wikipedia. (2017). Retrieved November 8, 2017 from https://en.wikipedia.org/wiki/Spotify

Wikipedia. 2018. Pearson correlation coefficient. (2018). Retrieved September 4, 2018 from https://en.wikipedia.org/wiki/Pearson_correlation_coefficient