Network Analysis of the Hip Hop Community

Kyle Hart
M.S. Candidate – Computer Science
University of Hawaii at Manoa

Abstract

In the early 2000's as hip hop was entering the popular zeitgeist there was an ongoing joke about how many artists would be featured on a single track. While collaborations were not unknown to other genres of music—occasionally 90's rockers would team up to form "supergroups" like Audioslave or the Foo Fighters—rappers took the practice to a whole new level. Eight or nine artists could be featured on a single track, a couple dozen might be credited to a single album. There were even some mercenaries with no career of their own own to speak of, but would gain fame by appearing on hundreds of others' tracks—perhaps the most notorious of these was *Lil Jon* who would show up on singles just to scream "Yeah!" a couple times.

If we see these artists and their collaborations as nodes and edges in a network we can run this graph through a gauntlet of metrics and models so that we might compare it to other commonly studied social networks and even to other genres of music. With this analysis we will delve into what makes this particular community unique and what processes are in play that contribute to this distinction

The Data

In November 2000 Portland based programmer Kevin Lewandowski launched a crowdsourced database called Discogs[1]. Currently this is the largest known public database of meta-data on commercial and non-commercial audio releases, containing information on nearly 10 million releases from over 5 million artists. The data can be accessed via a XML-based RESTful API or by direct download.

For this project we downloaded the most recent database release from April 2004 and extracted the desired data with a Python based SAX parser[2] .

Because the each release[3] is stored as an XML document with an entry listing the contributing artists, it was necessary to first construct a bipartite graph from from artists to releases, then to extract our collaboration network from the projection of this graph. Additionally, it should be noted that only a subset of the database was used, namely we were only interested in releases with 'Hip Hop' as one of their genre tags, and only those that had more than one artist credited. This means that it is possible, even likely, that there are some rappers left out that released only by themselves. The decision to leave out these isolates was motivated by the fact that we are interested in the *community*, which canonically disqualifies isolates in the same sense that backyard power generators are excluded from the analysis of a power grid.

Moreover, since many of the releases we credited to groups the decision was made to break these groups down into their actual members. This was necessary to get a full picture of the network due to the sheer frequency of which artists would work outside of their groups and due to the very nature of group productions. Unlike most pop or rock groups, a hip-hop group production has the same structure as one produced by multiple solo artists—each verse is given to an individual rapper, who will often write their lyrics independent of their collaborators input.

Finally, to get a genre for which we could run comparisons we went through the same process to derive a graph for 'Techno' artists. The motivation of this choice of genre is explained below.

Background

A career of a hip hop artist follows a fairly standard tract. As with other genres the young rapper generally starts off locally, preforming at

1 *Discogs: Vinyl Revolution* https://www.residentadvisor.net/features/1166

2 As with the R code written for this project, the Python scripts for data wrangling can be found in the following GitHub repository: https://github.com/gonzodeveloper/netsci_hiphop

3 A release is defined as a commercially or non-commercially released album or single, not an individual track

small shows independently or along side others long before they put out their first production. However, unlike other genres where these young artists are scouted by labels then signed into singles and album deals independently, budding rappers will often group up with other local MCs and DJs to increase their own notoriety before taking off on a solo career. Even in cases where local rappers are scouted by labels, they are often first debuted by recording tracks or verses on the albums of more prominent artists.

By contrast, techno artists follow a more traditional path. They start local, increase their following on the club-scene, then go on to release labeled production albums. Structurally the two genres could not be more different. While on a single track it is simple for a collection of MCs to alternate verses over a per-recorded *beat*, it is much more difficult for multiple techno DJs to contribute to same track. In fact, most collaborations for the latter that we are reviewing are over entire albums, rather then co-authorship of individual songs that we see from the former.

This domain information should help inform some of the basic metrics on the two network graphs and assist in the parameterization of our models.

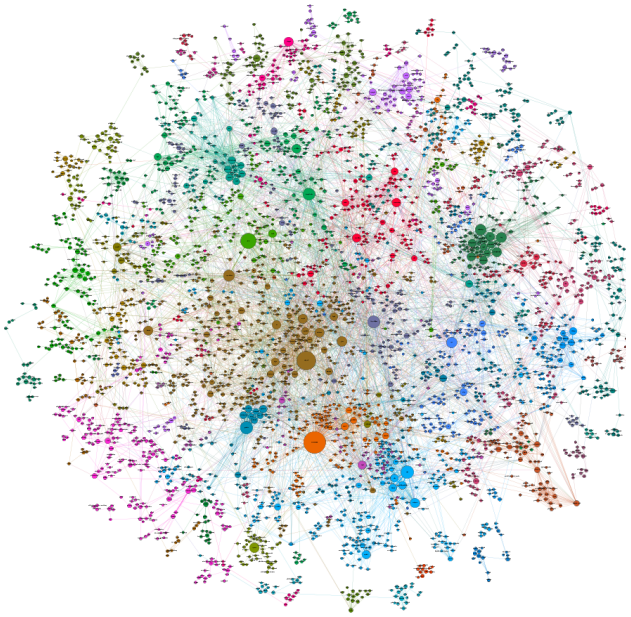


Illustration 1: Hip Hop Network- expanded with OpenOrd, colored by modularity class, sized by PageRank-- filtered for giant component

The Graphs

Metrics	***Hip Hop***	***Techno***
Vertices |V|	10,117	6,078
Edges |E|	13,845	5,688
Giant Comp. |V|	3,856	1,538
Giant Comp. |E|	9,301	2,169
Avg. Degree <*k*>	2.737	1.733
ln (|E|)	9.222	8,712
Mean Dist <*d*>	5.623	9.323
Diameter d_{max}	26	33
Cluster Coef $_{local}$	0.2344	0.3476
Cluster Coef $_{global}$	0.6366	0.5245
Degree Assortivity	0.2150	0.3402

It is immediately obvious that the techno artists' network gives us a far more sparse graph than that of the hip hop network. This naturally leads to a lower average degree as well as a higher mean distance and diameter, which is especially notable considering the techno network's smaller overall size. In fact, when compared to other graphs we find that the latter has basic metrics similar to the mobile-phone calling network[4]. However when we look at the hip hop network we can see that its relatively low average degree and short average path length more closely reflect that of the email network. These figures are not at all surprising. Given the differences of the two genres discussed earlier, we should have expected an overall higher level of collaboration in the hip hop network.

The difference in clustering coefficients (i.e. transitivity) in these graphs is perhaps most notable of all the basic metrics. That the rappers have a higher global coefficient is explained by their tendency to work with artists across the spectrum, with many even recording joint productions along side non-hip-hop names. This is a necessity for any artist hoping to broaden their appeal and grow their fan-base. It is surprising

4 Information on other well-studied networks can be found in Albert-László Barabási's textbook, *Network Science*

though, that the techno artists have a higher local figure. With the background discussion, our initial assumptions would have us predict that because rappers tend to work in groups they would have higher local transitivity. It seems though that there is enough collaboration with unconventional and unknown artists[5] in the hip hop graph that the local clustering coefficient is brought down. This in turn suggests that the hip hop network is quite open and susceptible to new ideas.

Finally, as one would expect with any real social network. Both of these graphs do fit firmly into the small world regime.

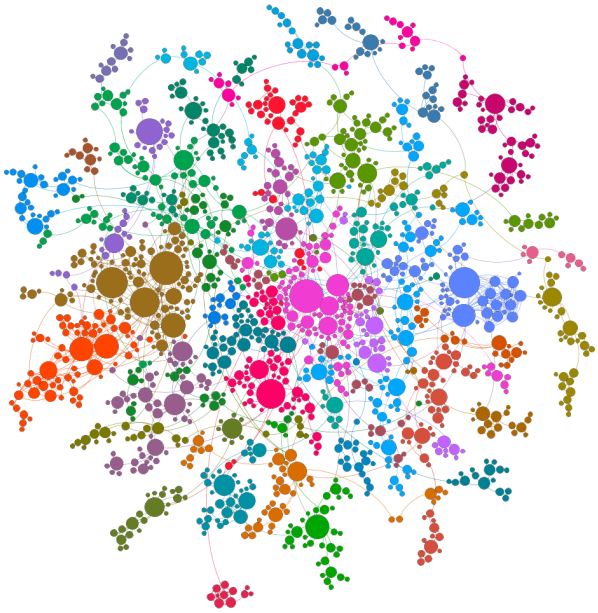


Illustration 2: Techno Network- expanded with OpenOrd, colored by modularity class, sized by PageRank-- filtered for giant component

Centrality

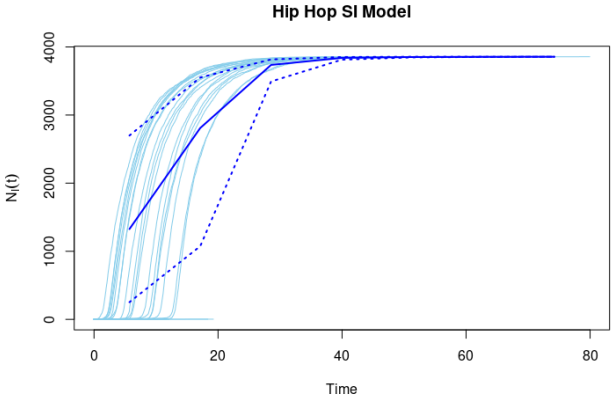
Since this study is written with the goal of examining the global structures of our networks we will not dedicate took much space to comparing the centrality metrics of individual artists in either graph. That the highest ranked artists by various centrality measures—such as degree, pagerank, betweeness and closeness--are indeed popular artists is unremarkable, unless the reader didn't trust the efficacy of traditional centrality metrics.

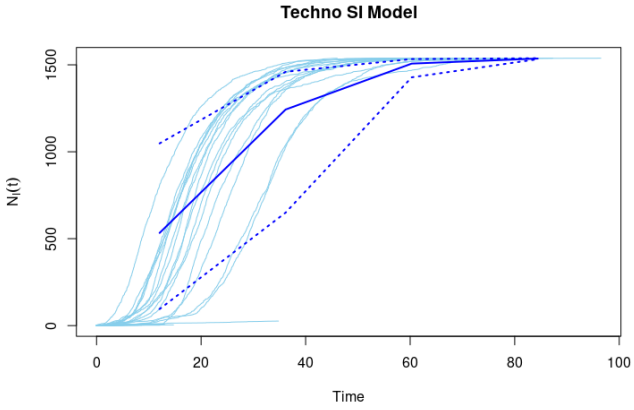
However, with one measure in particular, eigen centrality, we did have an interesting result. The top ten artists as scored by this metric were exactly the top ten members of the hip hop group *Wu Tang Clan*. Because we can simplify our understanding of eigen centrality to mean that *nodes will gain higher rank when connected to other nodes with high rank*, it is safe to conclude that in such a large graph, populated with many other such group ensembles, Wu Tang is particularly notable for its' members work with a spectacularly broad range of other artists.

To examine some of the other prominent artists we did construct their ego networks, though due to their size, it would be far more informative for the reader to explore their nodes in the full gephi visualization

Cascading Ideas and Failures

We can use the existing Susceptible-Infected (SI) model to explore how fast ideas might spread in both of these networks. For the model I chose the somewhat arbitrary *beta*—infection rate—to be 0.25, then ran 50 simulations on each graph to get the results.





5 The reader is left to explore the graph on their own for this trend. An prime example of this would be electronic artist Massive Attack's inclusion in the network because of a single collaboration with Mos Def

As predicted earlier, it is clear that ideas (or infections) spread much faster through the hip hop network. This is largely to to the high number of prominent artists with links spreading across the graph. Moreover, we can see the reflection of this in reality with the speed and pervasiveness of trends in rap music—DJ-808 drum machines, Auto-Tune, Mumble Rap, etc.

Furthermore, we can measure the robustness of these networks with a "targeted attack" simulation. Although, not a realistic scenario, we can test the robustness of the networks by removing high degree nodes in descending order and observe how the size of the giant component shrinks.

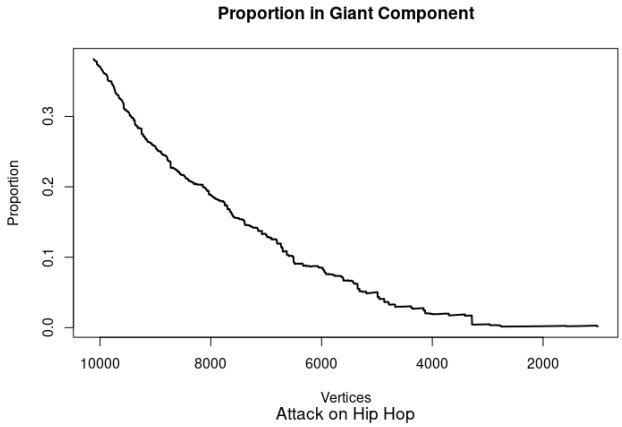


Illustration 3: Robustness simulation on Hip Hop Graph

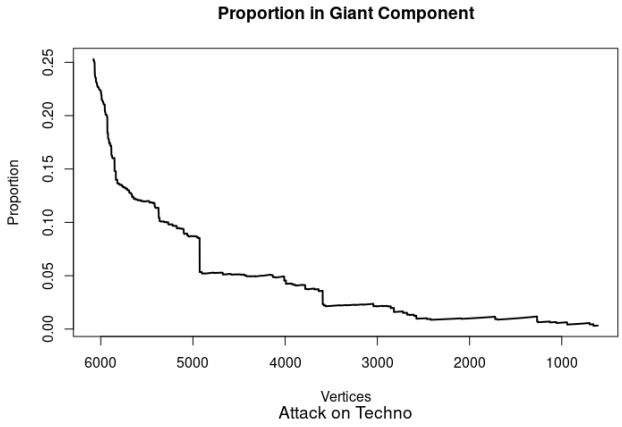


Illustration 4: Robustness simulation on Techno graph

The hip hop network shows amazing resilience to the attack. In fact, the decay of the giant component is not even particularly faster than if the attack were to be randomized[6]. On the other hand, the decay of the giant component when under attack in the techno graph reflects a that of a more typical scale-free network which relies on hubs to hold together its structure.

Modeling

It just so happens that the degree distributions for both of these networks does do not quite lend themselves to Barabasi's preferential attachment model. After a solid couple nights of fiddling with the various parameters the impossibility of properly fitting the models' geodesic distances, gamma values, and transitivity figures to the original degree distribution. This motivated the a somewhat more simplistic approach.

With both of these networks we see the existence of hubs along side a relatively low degree cutoff. This suggested the possibility of a stretched exponential or log -normal degree distribution.

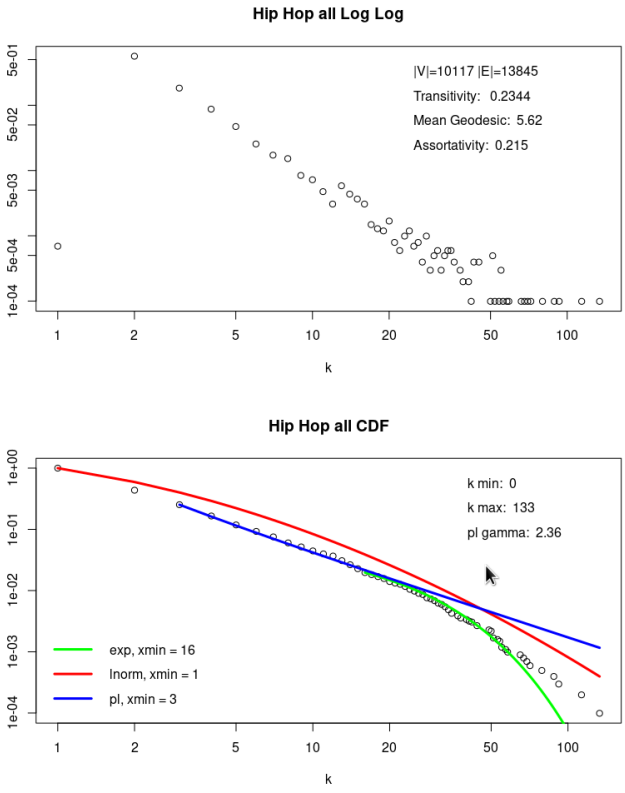


Illustration 5: Hip Hop Degree Distribution

6 See additional graphs in repository. As well as high resolution

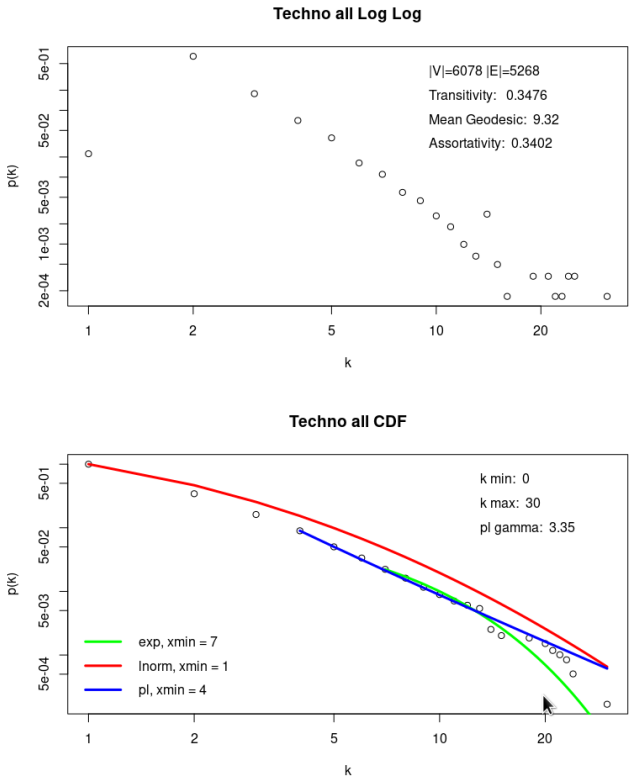


Illustration 6: Techno Degree Distribution

Informed by these plots we ran distribution comparison metrics on each, measuring fit of the log normal vs. the power law distribution. For both, though especially for the hip hop network, the test statistics strongly suggested that a log normal was a better fit.[7]

Considering the fact there is indeed an upper limit of how many artists can collaborate on a single release, and the reality that no artist, electronic or rapper can reasonably work with thousands of others over their career---as a fat tailed distribution might suggest—the stretched exponential does seem a sensible fit.

Unfortunately this leaves us without a model to explain other more pertinent graph metrics such as transitivity, distance, and assortitvity, so we will leave further discussion to the following sections.

Discussion

The metrics gathered on these graphs, particularly those relating to the hip hop network, do lead us to several conclusions, some more obvious than others.

In contrasting the two genres we did learn that each possesses a unique network structure. The differences in these structures can in turn be explained by the differences in the domains from which the emerged. Additionally, by exploring the background of each of the genres we were able to accurately predict certain graph metrics such as transitivity, average path length, and robustness.

The use of the SI model to show the potential spread of ideas within these genres did give us a for more novel insight. The model showed the high susceptibility of the hip hop network to new ideas, thereby confirming a somewhat long-shot assumption given earlier in the paper.

Further Study

The real success of this project is actually the construction of these networks so that they can be loaded and analyzed more in the future. There are many questions that still remain of which the author lacked the time and initiative to follow up on.

First and foremost, in the original proposal of this paper, we asked whether or not hip hop artists were more likely to collaborate based on region, age, or established success. These are all measures of assortitvity that only await the addition of the appropriate attributes for these graphs. However, because of the Discogs dataset did not include any of this information for their releases, such attributes would either have to be added manually or parsed from another database.

As mentioned in the last section, we also need to find a better model that could cover all of the various metrics of our networks in question. Perhaps the preferential attachment model could serve this purpose if more time was given to finding the correct parameterization, yet it seems more likely that the answer lies in the employment of exponential random graph models (ERGMs).

Finally, additional curation of the graphs themselves may be necessary. While the author did take the time to remove duplicates, merge aliases, and split up groups, because much of this was accomplished by hand, there likely still exist some minor inconsistencies in the data.

7 See main R script for exact figures