Contemporary Composers Web Archive

Nicole Greenhouse, Giulia Occhini, Pamela Graham
AU Datathon NY (in spirit)

(Attempted) Text Analysis

- Actually able to unpack the auk.tar.gz files using Powershell
- Realization that all of the Archives Unleashed tools use different softwares to use commandline--Anaconda to access Jupyter Notebooks; Gitbash to run grel
- Jupyter ended up being too high of a learning curve for me without direct in-person help, I couldn’t really figure out how to run the full text file in the Jupyter notebook (though I was able to run things a little bit yesterday in Google Colab before we ran out of RAM
- Was able to do some very basic GREL functions in Gitbash, but I think the grel was not working on the whole dataset given the results from the GREL (too few lines, only showing one website’s text
- General take-away, I’m glad I got to experiment with new tools, but for someone who has never done any data science activities, this was really hard to do without more in-person assistance

CCWA - networks

Getting started

After a day of failed data downloads, etc. etc. etc.

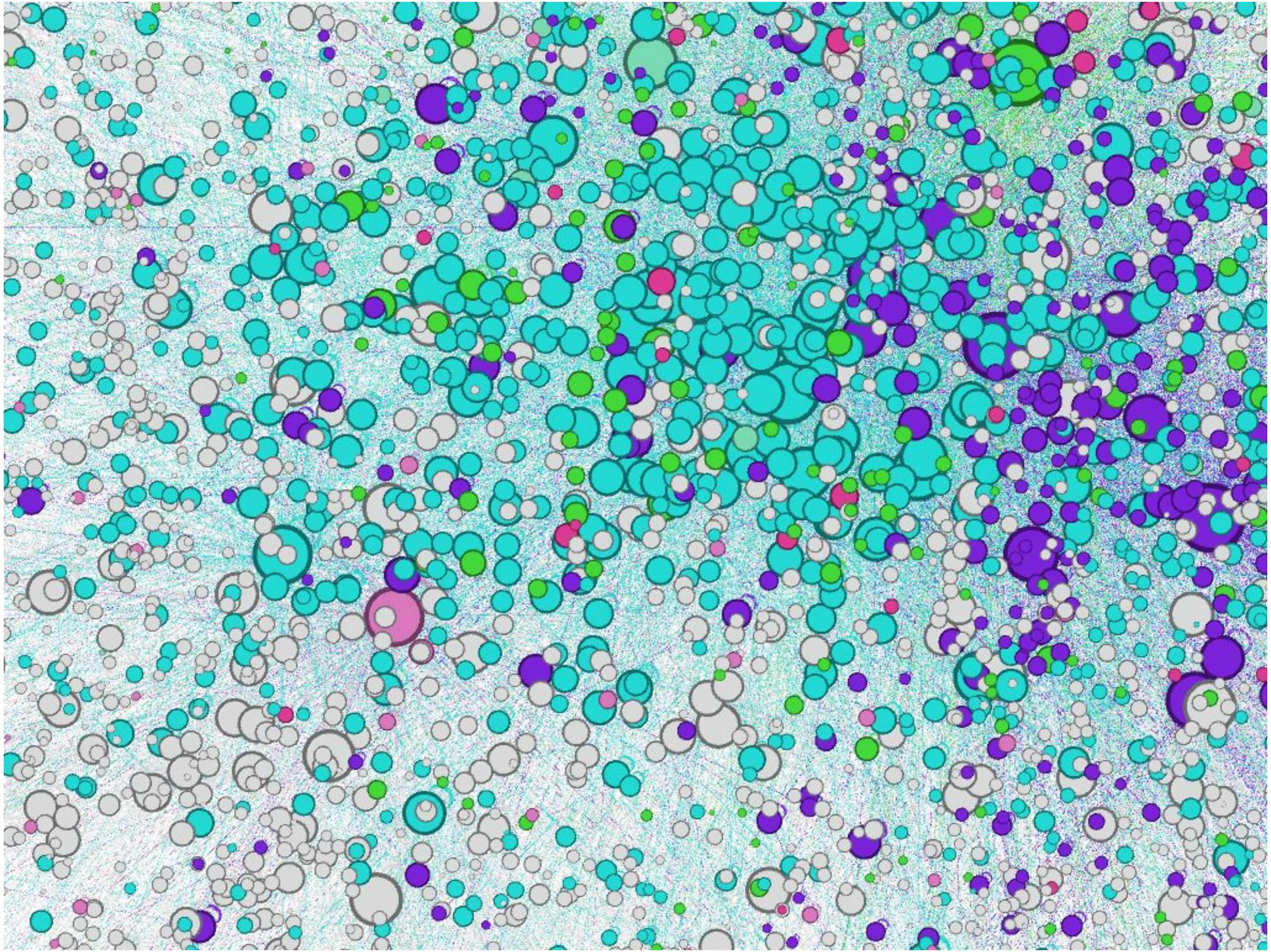
Special delivery of the CCWA gefx file got me going (thx ruebot!)

Note from Nicole: Also had trouble opening Gephi due to Java issues

Had downloaded Gephi in the past

Thanks to the quick Gephi Walkthrough that Ian gave us and Sarah's awesome tutorials

Imported the data. . .



Beautiful but meaningless.

What do I want to know?

Page Rank (most popular websites)

Modularity (patterns or clusters of community within a network)

Patterns of **In and Out linking**

Filters and Statistics

24,167 Nodes (individual websites)
34,477 Edges (connections between nodes) ***Correct me if I am wrong here!***

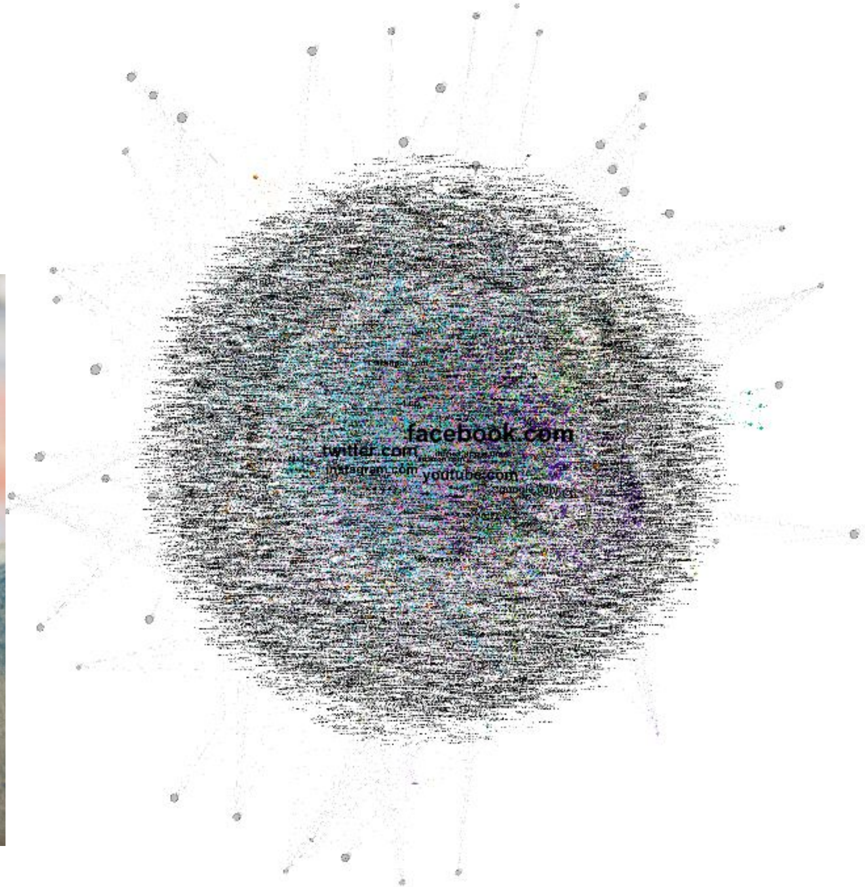
Filtered on Degree Range to limit to nodes with at least 10 links

690 Nodes AND 4,661 Edges

Ran modularity algorithm -- 53 communities in this collection

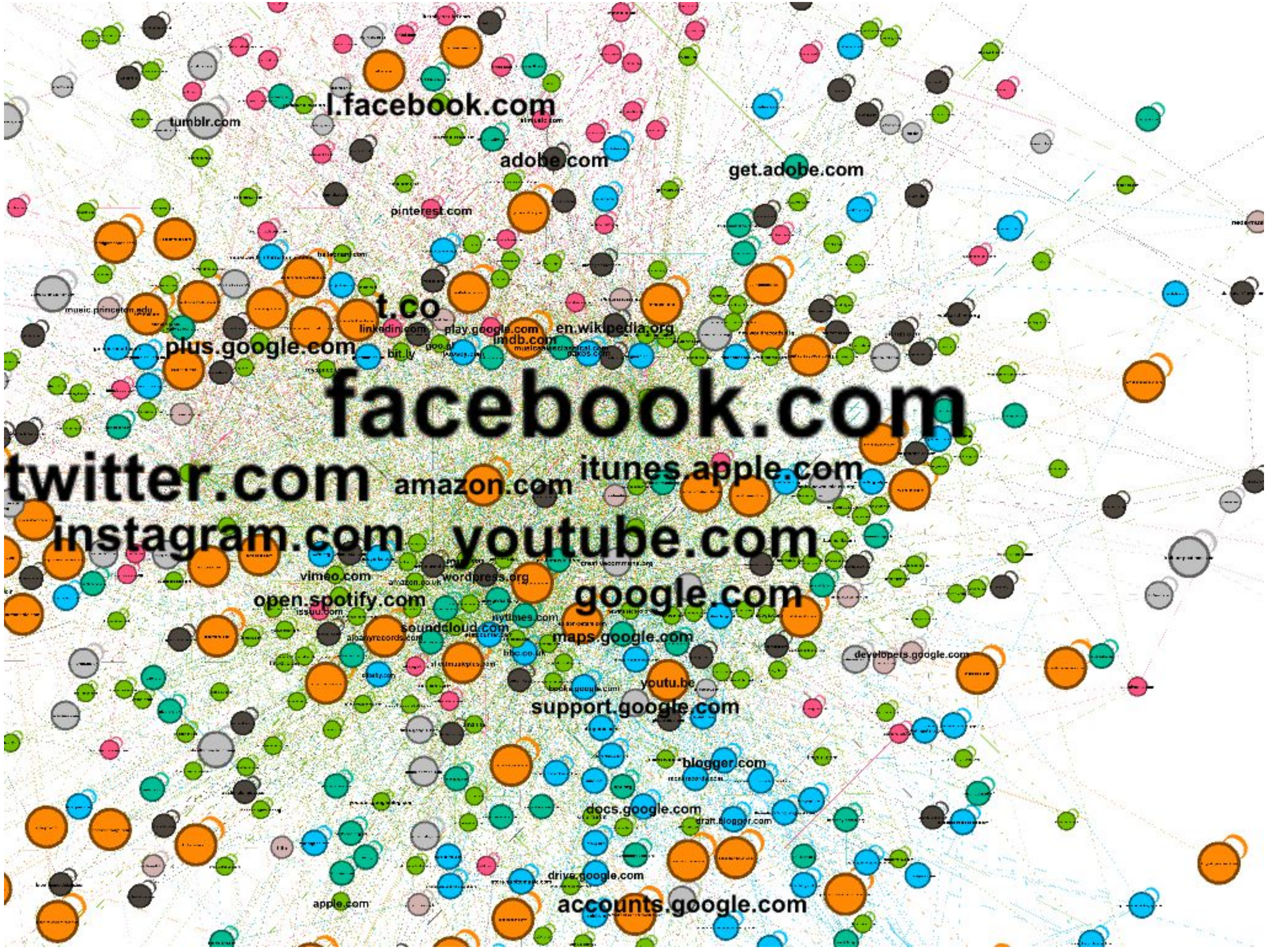
Ran connected components-- 34 weakly connected components and 557 strongly connected components

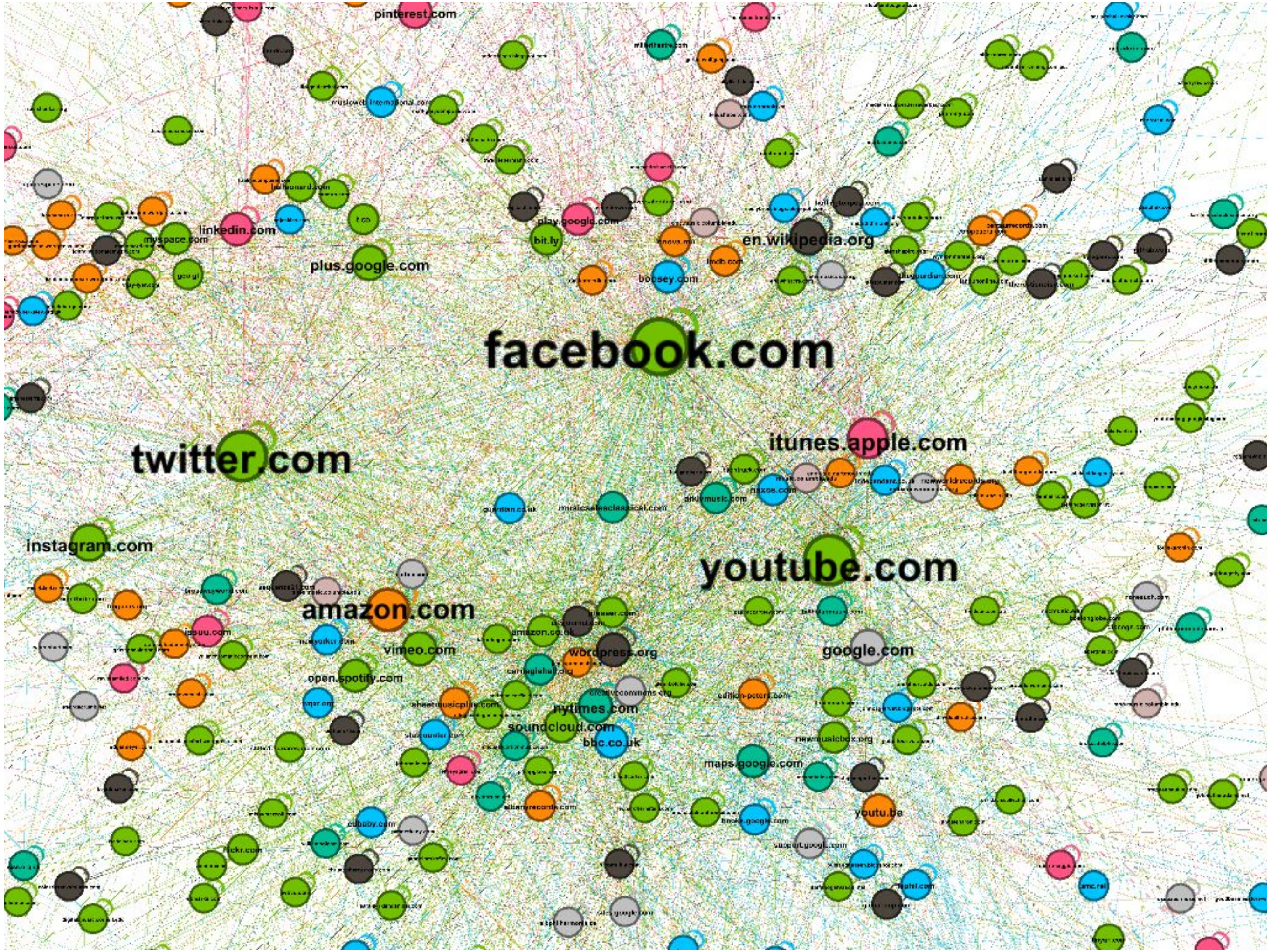
My first hairball



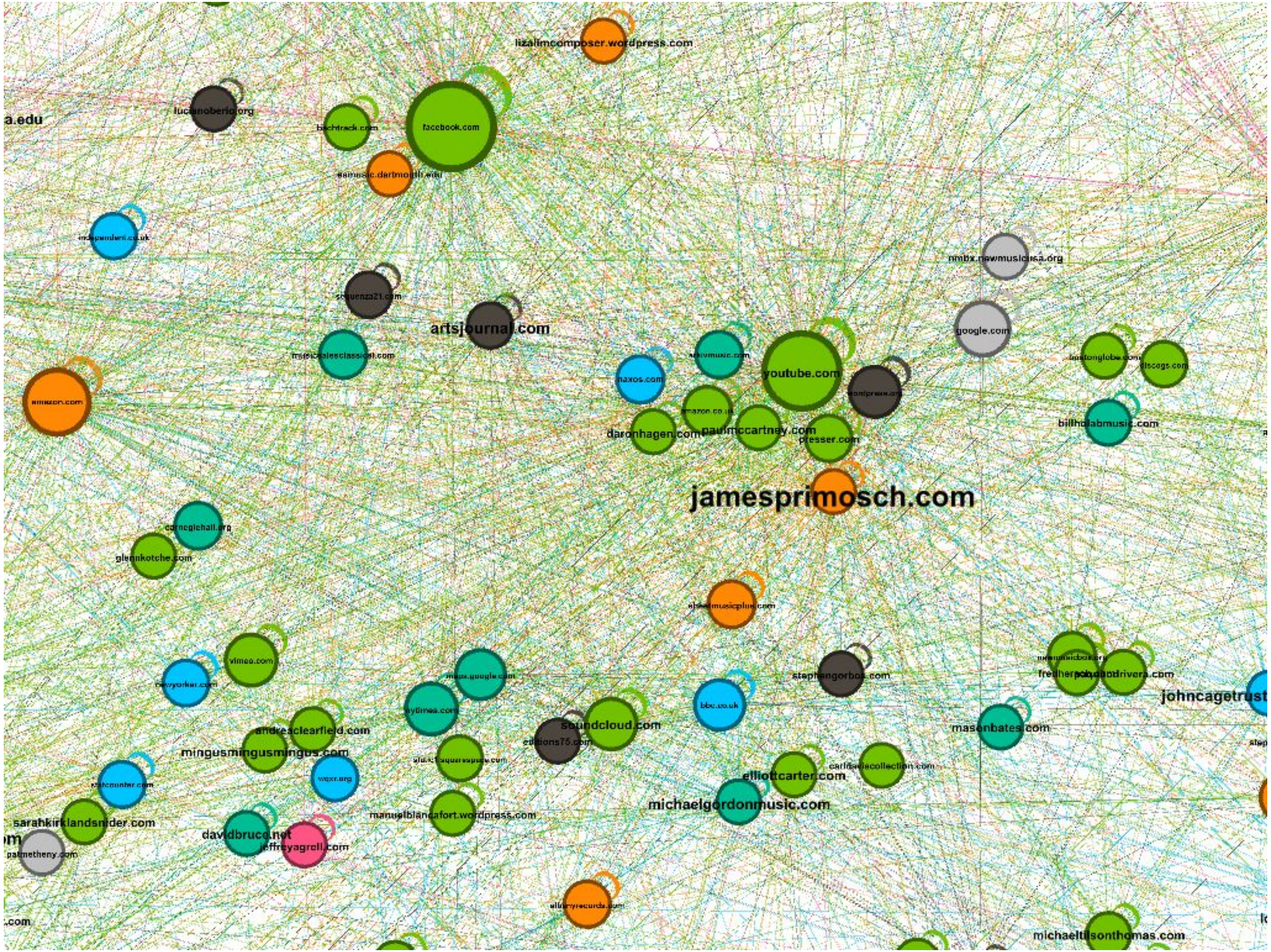
PageRank

Usual social media prominence





In links to nodes



Out links from nodes

James Primosch?

What did I learn from Gephi / link analysis?

Explore the 34 weakly connected components.

- What’s going on? Represent music that is more niche?
- Or is this helping us see areas that have been overlooked in our collecting?

Learn more about the 53 communities

- Is this telling us anything about how different composers connect? What attributes, factors could explain such possible connections?

Explore the dominant nodes

- No surprise to see social media; more to learn about relative strengths?

What else did our group learn?

Data wrangling is really hard. Lots of environments, platforms, etc. etc.

Big learning curve if you aren’t already familiar with a lot of this

Mastering DH/analytics tools is pretty crucial

Hard to come up with questions until you have played around a bit with data

What else did our group learn? [Giulia]

Learned more about PySpark: I had some training in Spark in the past and I hated it profoundly, now I am a bit more positive towards it

Learned that Google Colab is not the infinite resource I thought it was :confused:

Reasoned a lot on my name+surname problem

Discover that my text analytics skills are actually not that bad

Looked inside a lot of very cool archive material! I was really positively surprised by the variety of content we could chose from

MOST IMPORTANT POINT: Discovered that a lot of other people are interested in representation, which is a topic that I have deeply at heart