Effects of Decision-Making when Generating a Network from Soundcloud

BY: BUXTON, JOSIAH

GODLEY, CHRISTOPHER

Why are we scraping data?

Our network analysis is computationally intensive and would take too long to finish on the full network. (Clauset, Lecture9 2017)

Soundcloud does not have a freely available API for attaining user data.

Why Soundcloud?

- We're both musicians and were naturally curious
- Increasing commercial potential of Soundcloud artists
- Profile pages provided an easy interface for html based scraping

Big Goals

- Top Artist Prediction
- Link Prediction

The Woes of Scraping Data

Loading webpages
- Some artists have over 1 million followers with unique urls.
- Unreasonable to gather all data
- DDoS protection prevents from scraping too fast

Scrolling through web pages
- Soundcloud only loads a small amount of artists to display
- Used a package to trigger a scroll event to load more of our desired data

Design Decisions
- How many artists to grab?
- How to branch to new artists?
- What attributes to grab?

Network Topologies: Planning

Types of sampling to be used:

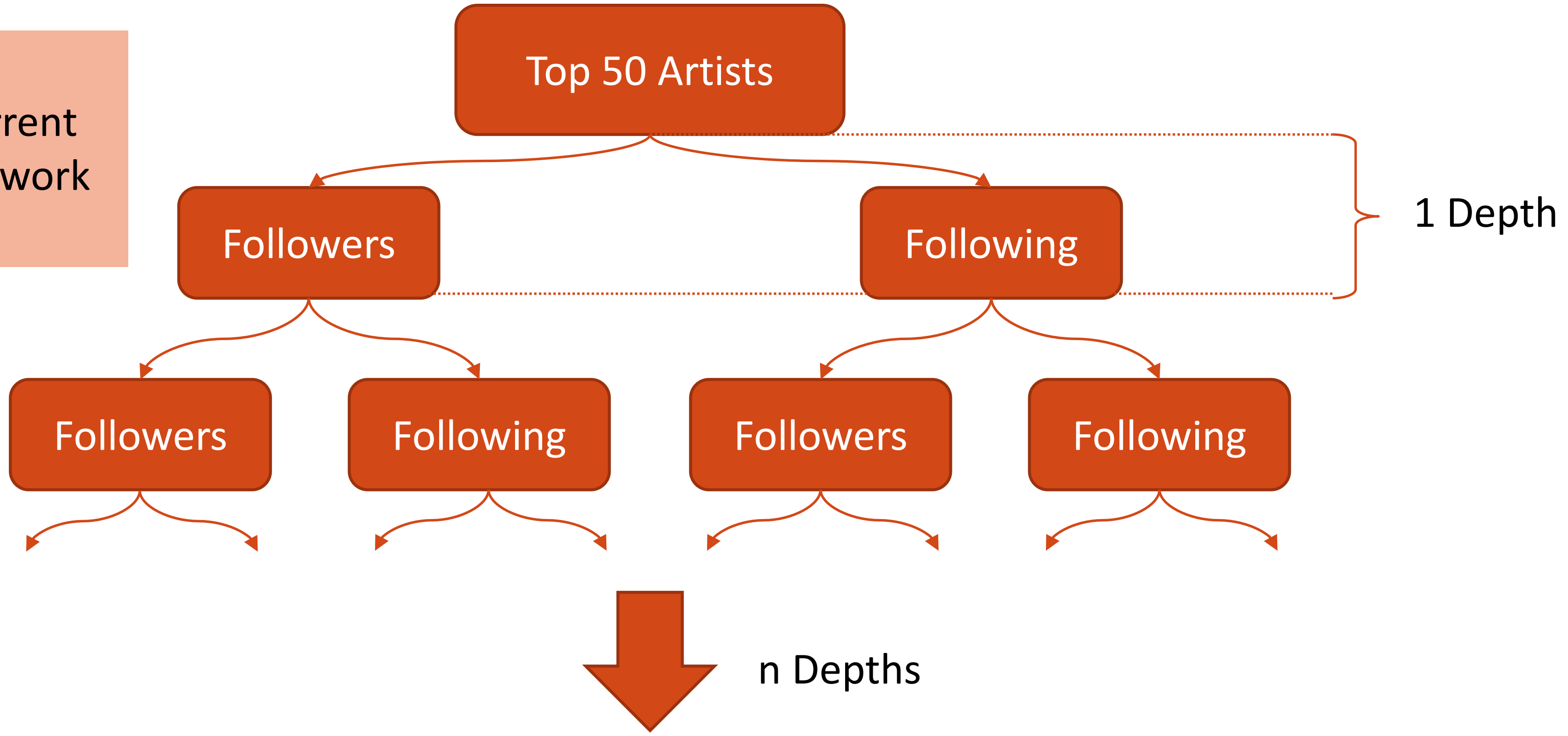
- Snowball Sampling
  - for each seed vertex i, and distance *l,* include all vertices (and their neighbors) for an *l*-step breadth-first search tree rooted at i
  - In our models, *l* = depth, and i = a particular artist
  - The snowballing can occur via links to an artists FOLLOWERS or FOLLOWING accounts
- Adaptive Sampling
  - for each seed vertex i, and integer s, include all vertices (and their neighbors), or include all edges, in an adaptively-grown tree containing s vertices rooted at i

- Different approaches to adaptively sampling the Soundcloud network:
  - Removing followers
  - Only adding artists (number of tracks > 0)
  - Only adding artists/users over a certain number of followers

Network Topologies: Take 1

Depth = 0:
Each artist from the current top 50 songs is added as network nodes.

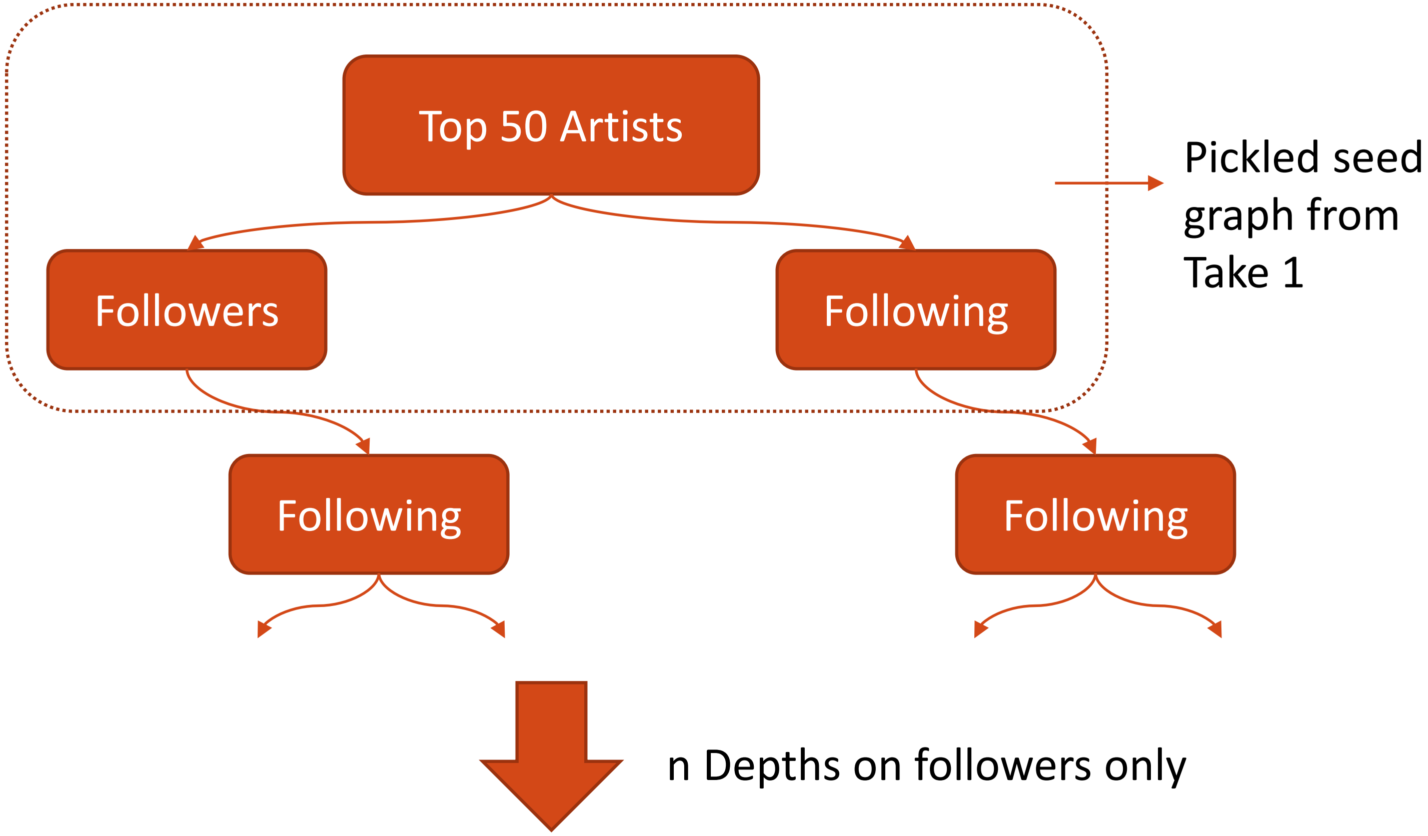
Depth > 0:
Each artist's "following" list is added as new artists, as well as a random sample of followers.



Network Topologies: Take 2

Depth = 0:
Each artist from the current top 50 songs is added as network nodes.

Depth > 0:
Each artist's "following" list is added as new artists. The random sample of followers is added to the network but not branched.



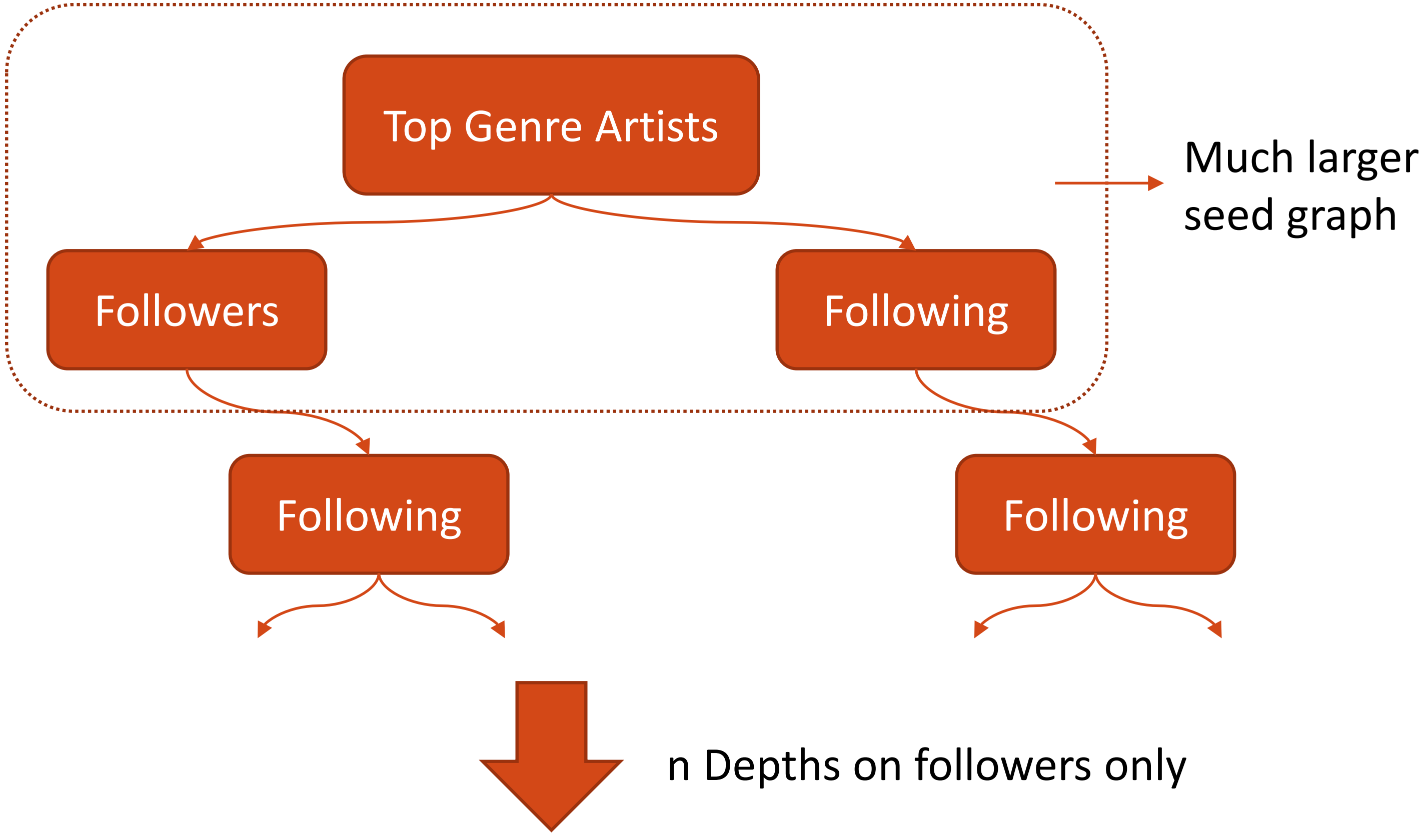
Network Topologies: Take 3

Depth = 0:

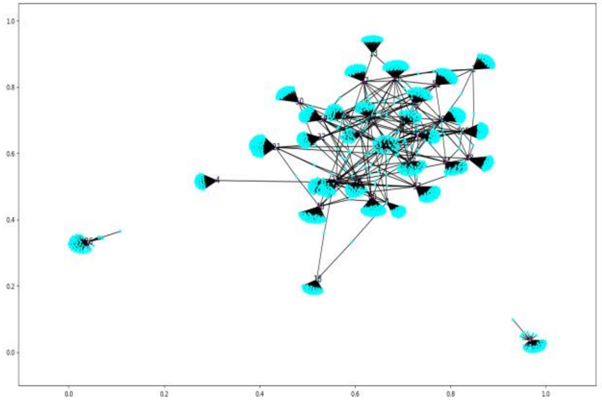
Each artist from the current top 50 songs of the top 30 genres is added as network nodes.

Depth > 0:

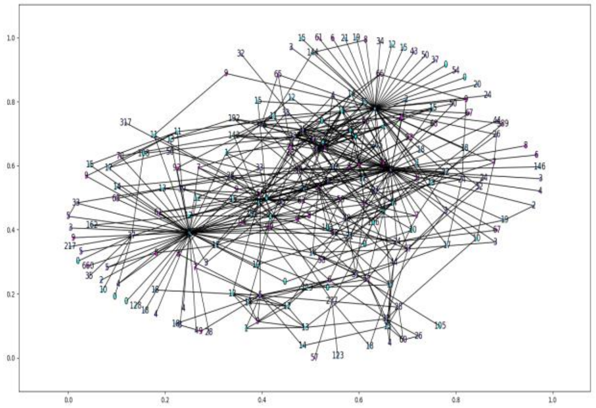
Each artist's "following" list is added as new artists. The random sample of followers is added to the network but not branched.



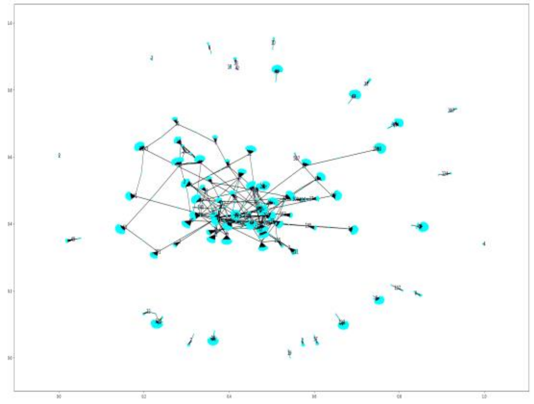
Network Topologies: Visualized



Take 1
N = 21258
Edges = 29210



Take 2
N = 39734
Edges = 59836



Take 3
N = 5598
Edges = 6016

Metrics

Clustering Coefficient (Nodes)

- Fraction of possible triangles through the node that exist

Number of Triangles

Degree

GbA Heuristic

- Guessing attributes such as whether a node is a user or an artist

Link Prediction

- Degree Product AUC
- Common Neighbors AUC
- Shortest Path AUC

Graph Metrics Detailed

	Take 1	Take 2	Take 3
N	21258	39734	5598
Edges	29210	59836	6016
Number of Triangles	11715	11229	411
Max Clustering Coeff	1 (The Actual Tanis)	1 (SEBASTIAN)	1 (MAX)
Clustering Coeff	0.023	0.012	0.0095
Max Degree	183 (Wicca Phase GBC ETERNAL, 25k followers)	184 (G-EAZY, 1.4million followers)	157 (IOF,
Mean Degree	2.75	3.02	2.15

Future Work:

Sampling Decisions:

- Add larger sampling of followers at every step and connect any existing artists they follow
  - Generate a larger network in general
- Probabilistic Sampling
  - Add edges with a probability, to lessen possible bias applied when sampling from only the popular artists

Random Graphs:

- Comparing our model to random graph models such as Erdos-Renyi, configuration model, etc.

Metrics:

- Link Prediction
  - Need larger sampling of followers (user)