Seed Scientific Data Engineering Applicant Assignment

Melissa Loudon
1 June 2015

Goal

Calculate tag similarity for the top tags applied to the top 1000 artists on last.fm. The solution described below also calculates artist similarity, and generates some network plots for artist recommendations by tag.

Code

In the attached code, extract.py contains methods to extract a list of top artists, then the top tags for each artist, from the last.fm API.The API is a bit temperamental, with requests to the top tags method occasionally failing or returning a response with no tag data. To mitigate this, the list of artists for whom tag data has been saved is checked against the full list, and the top tags call is retried for missing artists. The retries setting in settings.py controls the number of retries.

As tag data is obtained for each artist, a record with the artist name as key and a list of tag names as value is written to Redis as a list type. Tag names are stored directly; in practice a tag id with a name to id mapping would be more efficient. Redis is a good choice for a dataset that can fit in memory. A relational or nosql database would be better for a larger dataset and/or a more complex data model.

I'm using networkx to calculate similarity given by the Jaccard coefficient - the number of common neighbours of nodes a and b, divided by the number of possible common neighbours of the two nodes. In network_calcs.py there are methods to load the tag data into a bipartite network of tags and artists, then calculate similarity for a given set of pairs. Similarity is only calculated for nodes that have at least one common neighbour. I don't actually need to treat the artist-tag dataset as a network to calculate similarity based on immediate neighbours, but networkx handles two-mode networks neatly. If the dataset was too big to fit in memory, the Jaccard coefficient could be calculated from adjacency lists stored on disk.

Similarity data for artists and tags, as well as an intermediate dataset giving the artist-tag adjacency list, is written to csv files and attached as output. A real solution would store this data so that it is accessible by key. While the data fits in memory a Redis scored set would be a fast way to do this, assuming the main use-case is finding the most similar artists/tags.

Once I had the data, I wanted to visualize how it might look as a recommendation system. I wrote a function to get all artists for a given tag, and their five most similar artists, and write this to a csv file. I then used R to make 'recommendation' network diagrams for a few example tags.

Results

The artist-tag network has 17 436 nodes (1000 artists, 16 436 tags) and 93 329 edges. There are 495 996 artist pairs with at least one tag in common, of a possible 499 500 pairs. There are 1 913 910 tag pairs with at least one artist in common, of a possible 135 062 830 tag pairs.

A total of 101 159 tag pairs have a perfect similarity score of 1, and most likely occur only once in the dataset. Depending on the use case, a weighted similarity measure might either penalize or reward infrequently occurring tags. Because last.fm tags are user-defined, it might be better to just remove infrequently occuring tags from the dataset.

The 20 most similar artists are shown in Table 1. On the next several pages, I've included my 'recommendation' network plots - for a given tag, these show all artists with that tag, as well as their five most similar artists.

Table 1: Most similar artists

Artist 1	***Artist 2***	***Similarity***
Damien Jurado	Devendra Banhart	0.7094017094
Devendra Banhart	Beirut	0.7094017094
Damien Jurado	Beirut	0.6806722689
Damien Jurado	Sufjan Stevens	0.6666666667
The Jimi Hendrix Experience	Jimi Hendrix	0.652892562
Beirut	Sufjan Stevens	0.6393442623
Devendra Banhart	Sufjan Stevens	0.6393442623
Elliott Smith	Sufjan Stevens	0.6393442623
Iron & Wine	Sufjan Stevens	0.6393442623
Keane	Coldplay	0.6129032258
Bob Marley & The Wailers	Bob Marley	0.6
Damien Jurado	Elliott Smith	0.6
Keane	Travis	0.6
Damien Jurado	Iron & Wine	0.5873015873
Franz Ferdinand	Bloc Party	0.5873015873
Iron & Wine	Nick Drake	0.5873015873
Jennifer Lopez	Rihanna	0.5873015873
Portishead	Massive Attack	0.5873015873
Fatboy Slim	The Chemical Brothers	0.5748031496
Korn	Slipknot	0.5748031496

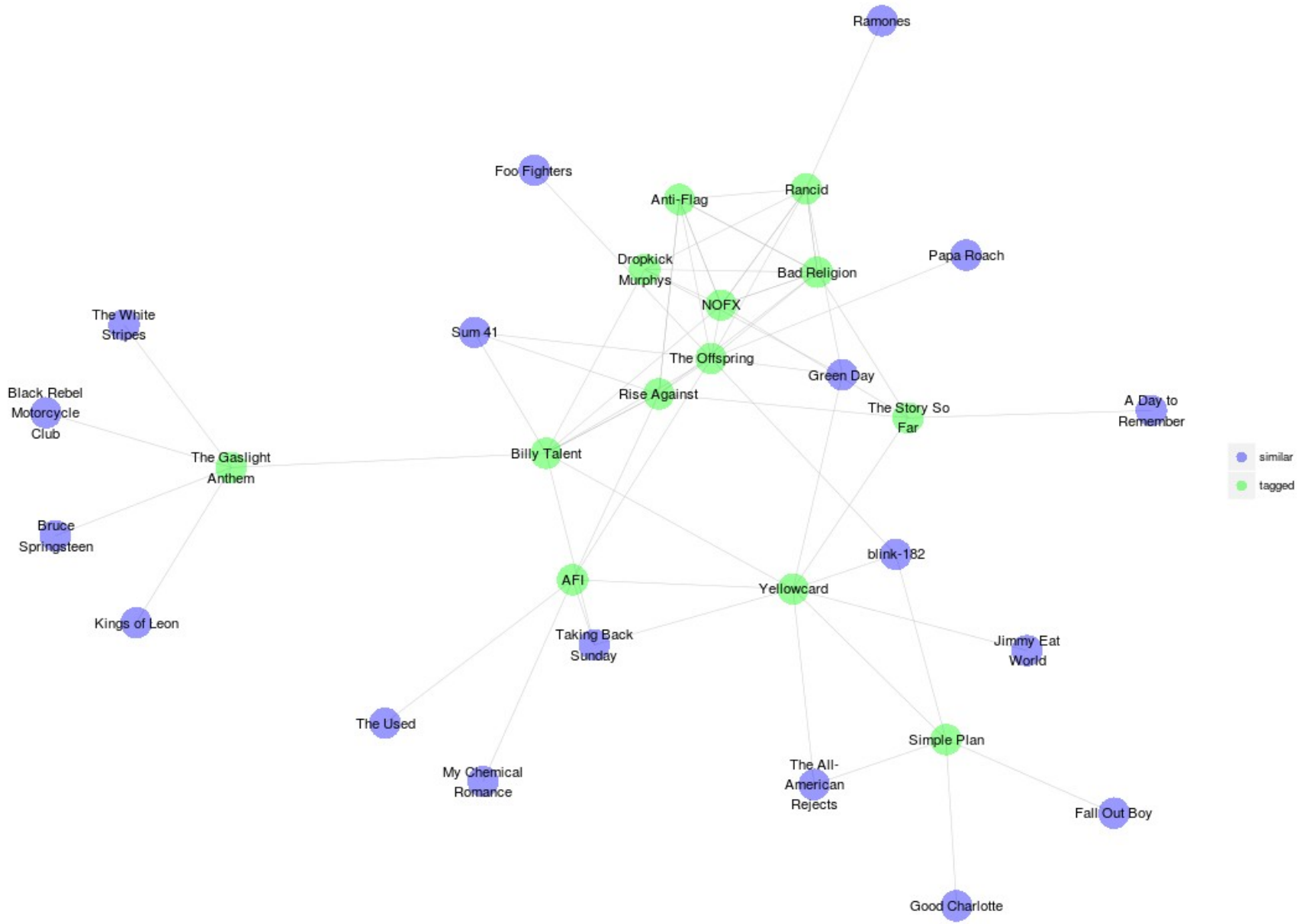

Figure 1: Melodic Punk

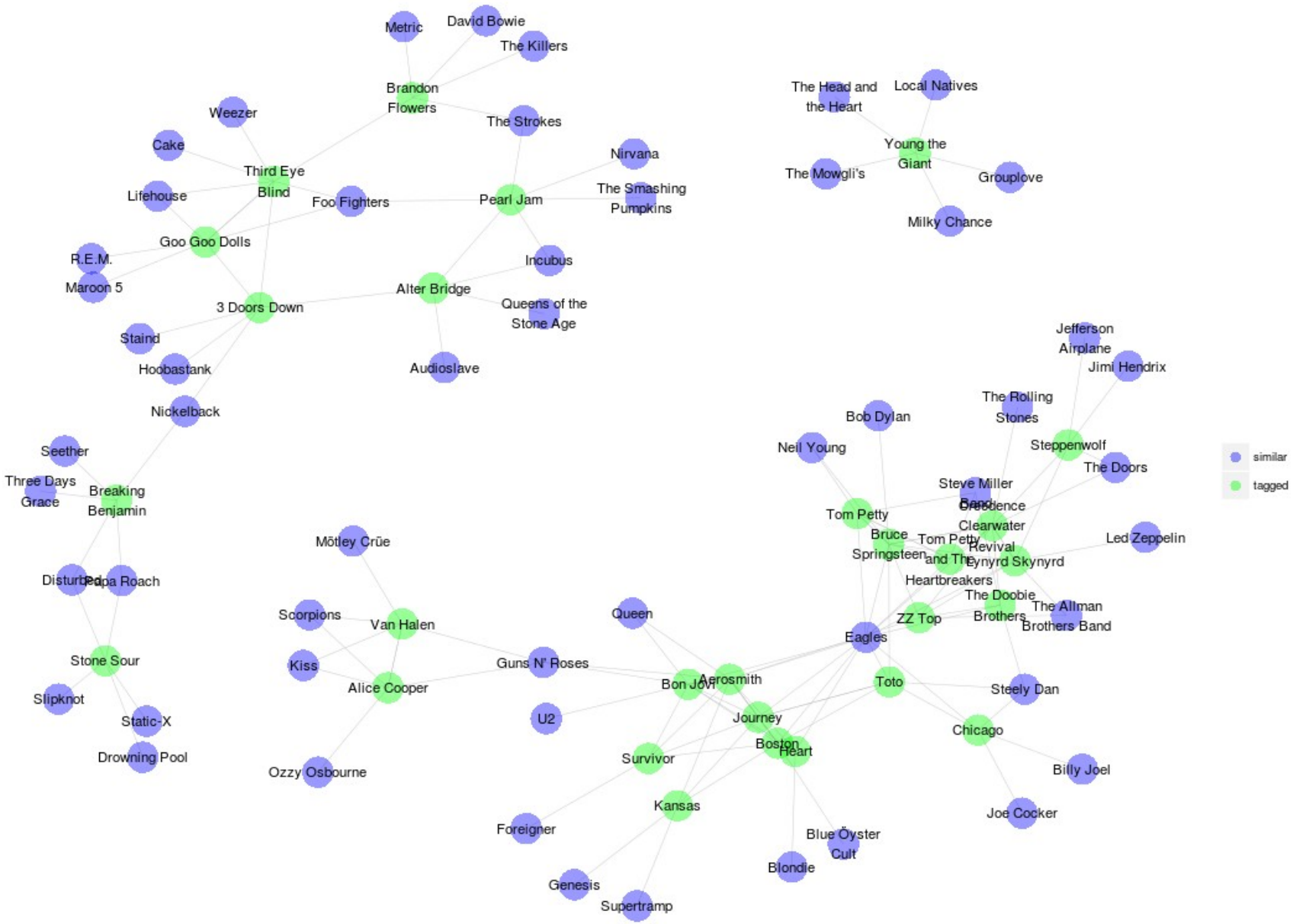

Figure 2: American Rock

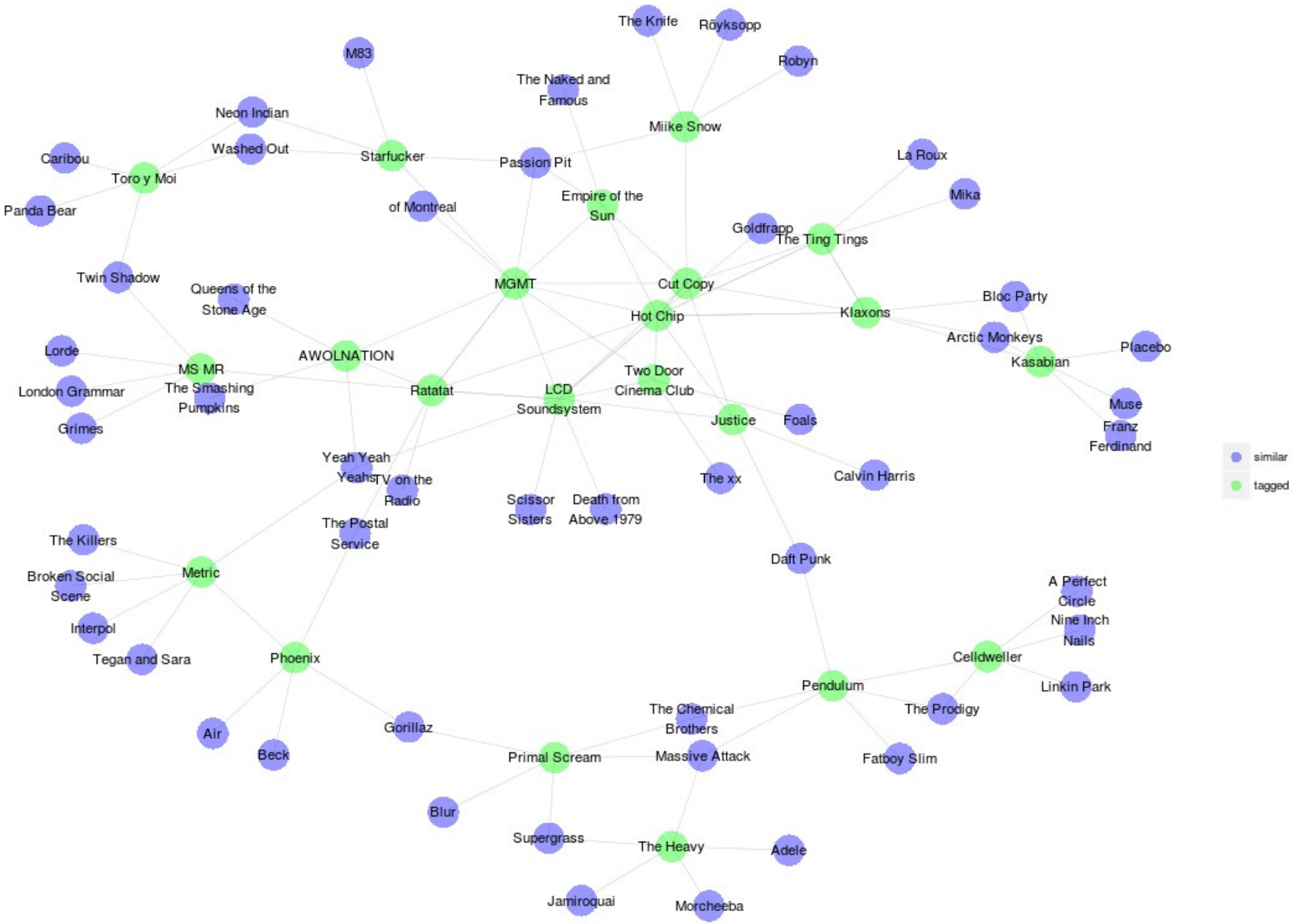

Figure 3: Electro Rock

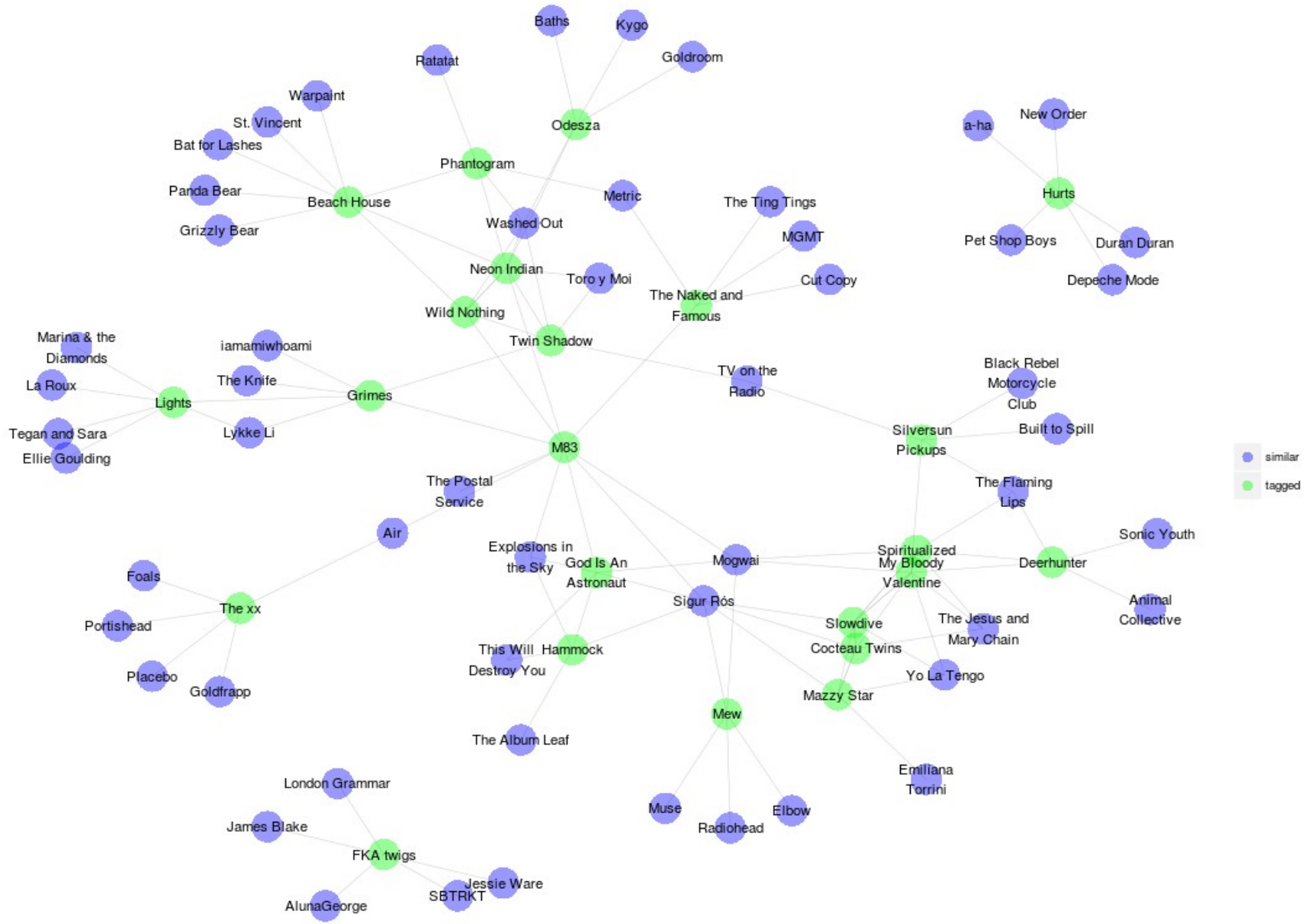

Figure 4: Dreampop

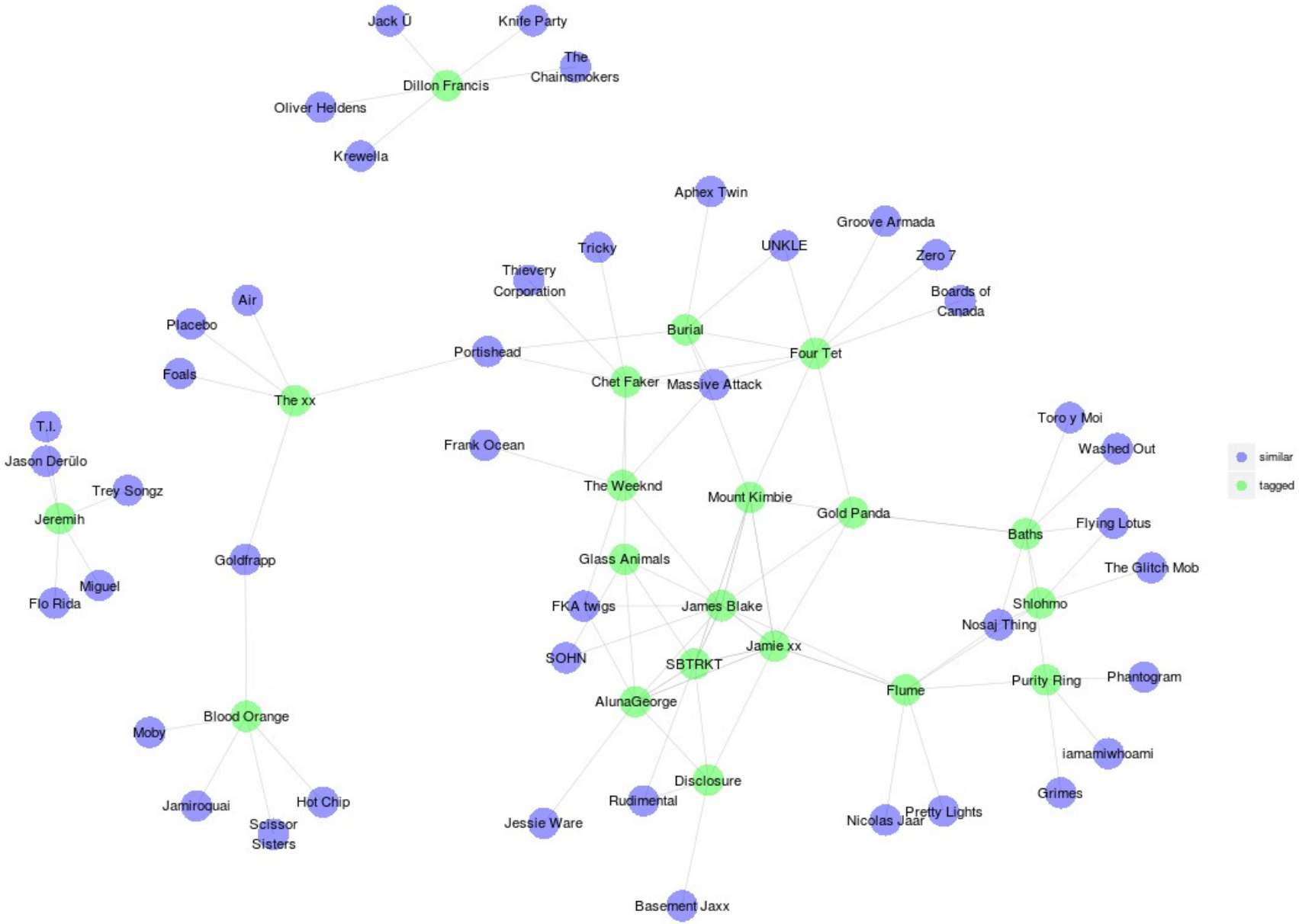

Figure 5: Post-Dubstep