Social Network Analysis of Music Communities in Major Cities

A direct comparison of communities of active music fans in Glasgow and Barcelona

Dewi Gould

Under the supervision of Drs. Luce Prignano and Emanuele Cozzo

26th June - 28th July 2017

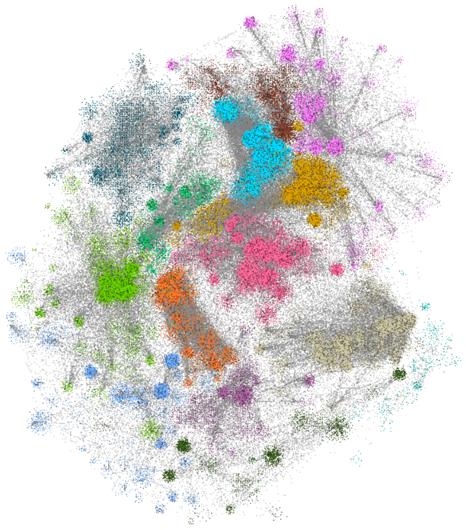

Figure 1: Network of music fans and their social media interactions with the most popular bands in Glasgow. Colours represent regions associated with specific bands.

Abstract

Basic social network analysis and graph theory techniques were used to examine and compare the community structure of music fans in two large and diverse European cities. Data collected from the public Facebook pages of popular bands was used to generate a network of around 100,000 Facebook users and the posts with which they engage. Individual modules and their roles within the larger community were identified for both Glasgow and Barcelona. A comparison of the two showed that music fans in Glasgow form a more isolated and insular community, whereas those in Barcelona are much better connected. Analysis on the composition of these modules led to some interesting findings, such as the influence of the 'type' of post (status, video etc.) and the mathematical description of the user engagement per post. These areas represent regions of interest, but require further exploration.

Introduction

The ‘music scene’ of a major city is clearly a very subjective notion. In this project, I wanted to find the music which makes a given city distinctive from others in order to properly analyse the behaviour of a city’s ‘true’ music fans. Initial work using bands found by searching for ‘the most popular bands’ in a given city yielded an enormous choice, depending on age and genre preference of a person. Spotify’s Music Insights website[1] provided the perfect solution to this problem. It generates a list of the music unique to each city in the World (within reason), comprised of songs which makes said city stand out from all others.

This project is centred upon the analysis of social networks, and as such the Facebook pages of these bands were the primary source of data. By looking at the last 50 posts of each band, and how fans engage with these posts, it was possible to generate a visual ‘map’ of the city’s music community (see *Fig. 1*).

The Facebook data was acquired using Netvizz v1.44 - a public app available to all Facebook users. Many studies have shown that this form of data representation in conjunction with graph theory can yield remarkable results: examining a collection of data from different sources can lead to very interesting and otherwise un-obtainable results.

Theoretical Background and Method

In this case, a bipartite network was formed with nodes representing posts by a Facebook page and the users who interact with said posts. Edges were formed if a user interacted with a post. For the analysis in this project, it made computational and logical sense to use the projection of this graph onto only the post nodes. This meant that the hidden communities within the larger network could be visualized. In this format, edges were formed between two posts if they shared a common Facebook user. The weight of an edge, w_{ij}, between two nodes i and j , was calculated using the same method detailed by Newman[2] in his paper on scientific publishing. It is variable depending on the number of common Facebook users and how many posts a given user k had engaged with, n_k.

$$w_{ij} = \sum_k \frac{\delta_i^k \delta_j^k}{n_k - 1} \tag{1}$$

Kronecker Deltas δ_i^k equal 1 if user k engages with post i and 0 otherwise, and the whole expression is summed over all users. This way, posts connected by many common users are given more weight, but if a user has engaged with many posts their contribution is diminished.

Netvizz yields .gdf files which can be analysed using Python’s NetworkX module and the network software Gephi. The biggest challenge of this project was having to handle data in various file formats and being able to convert between these formats for various different pieces of analysis. Different Python codes were written to handle these changes and compute the required information.

The most informative analysis performed in this project was looking at the community structure of the networks. Modularity is one measure of the structure of networks or graphs. It was designed to measure the strength of division of a network into modules (or communities). In order to partition the graph into these modules, the Louvain method was implemented through Gephi’s modularity function.

The underlying statistics of this were then analysed using the ‘role based description of complex networks’ detailed by Guimera, Amaral and Sales-Pardo[3][4]. This involved calculating the ‘withinin-module degree’, z_i, and the ‘participation coefficient’, P_i. The $z-score$ measures how ‘well-connected’ a node is to all the others in its module, whereas the participation coefficient defines how the nodes links are distributed among the other modules. This particular method allows for nodes to be split into various ‘roles’ depending on their individual values, a feature which is utilised in this project by plotting the nodes in the $z - P$ plane.

Results and Discussion

To give an appreciation for the macroscopic features of the networks, the projected graphs were plotted on Gephi (see *Fig.2*). It is clear that the Barcelona graph is far more crowded and densely packed than Glasgow’s: an early indication towards the idea that Barcelona’s music community is significantly better connected that Glasgow’s.

(a) Graph of Glasgow network projected onto posts. Different colours are associated with posts of different bands

(b) Barcelona network projected onto posts. Different colours are associated with posts of different bands

Figure 2: Projected graphs for Glasgow and Barcelona

By plotting the data in the aforementioned $z - P$ plane, it was possible to examine how the nodes were distributed in terms of the community (see *Fig.3*).

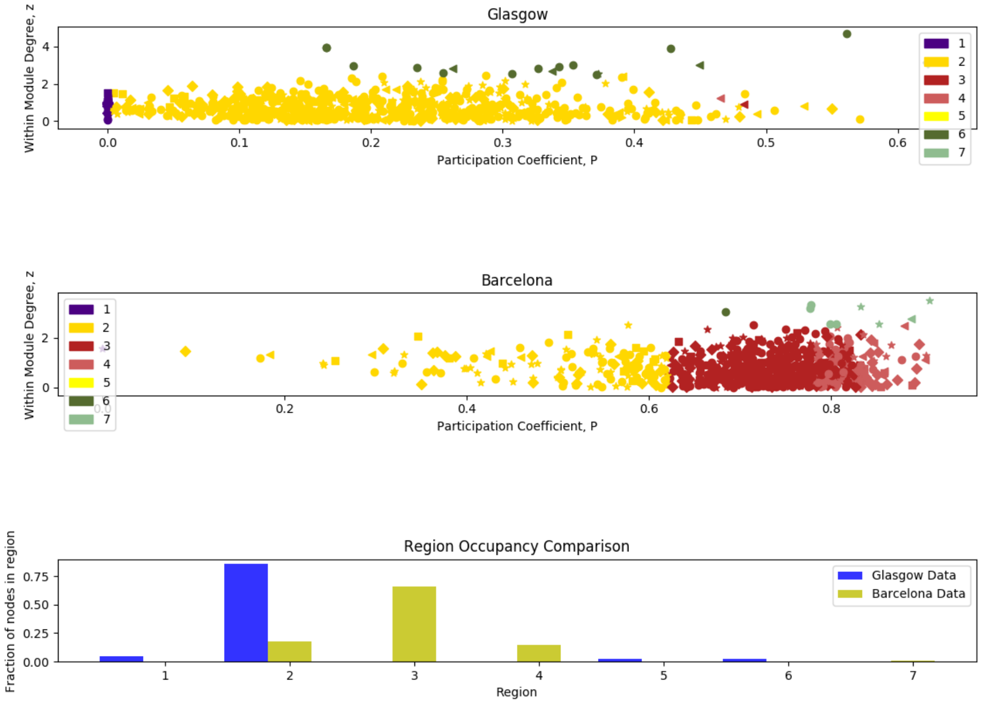

Figure 3: z-P plots for Glasgow and Barcelona. Colours represent different roles, and markers represent different types of post.

There is clearly some noticeable differences between the two cities. The majority of Glasgow's posts lie within Region 2 (known as 'Peripheral Nodes'): these are nodes with at least 60% of their links within the module. In comparison, the majority of Barcelona's posts lie within Region 3 (known as 'Non-Hub Connectors'): nodes with half of its links within the module. However, there are also a significant number of Barcelona's nodes in Regions above 2 and 3, unlike in the case for Glasgow. This is indicative of the fact that most of the posts in Glasgow are poorly connected outside of their own cliques, whereas music fans and bands in Barcelona are more likely to be globally connected. This difference is more starkly highlighted in *Fig.4*.

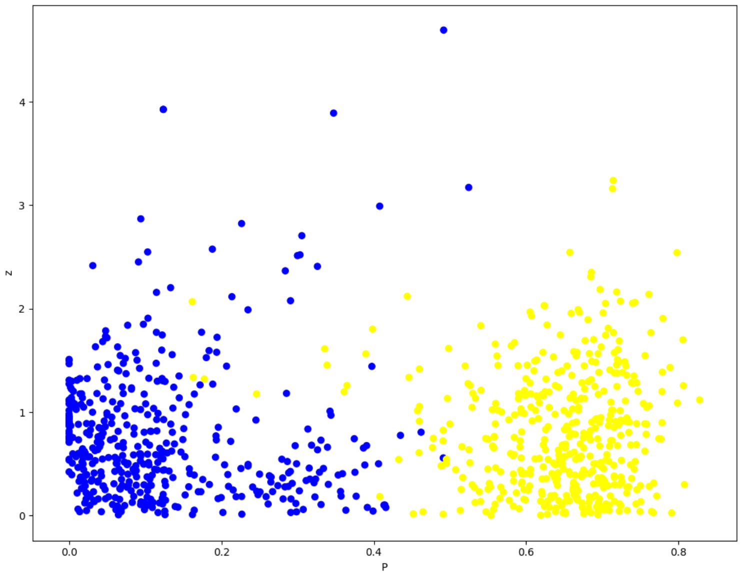

Figure 4: z-P plot for Glasgow and Barcelona combined (blue data for Glasgow, and yellow for Barcelona).

This shows that the data for Glasgow and Barcelona are almost entirely separated and distinct, a rather surprising result. Further analysis was performed on the composition of these regions. *Fig.5* shows the spread of the nodes among the different types of post. The chart shows the fraction of nodes in each region that are of a certain type. This form of comparison could be very useful, as finding out which type of post gathers the most traction would of course be extremely useful to the Facebook page administrators.

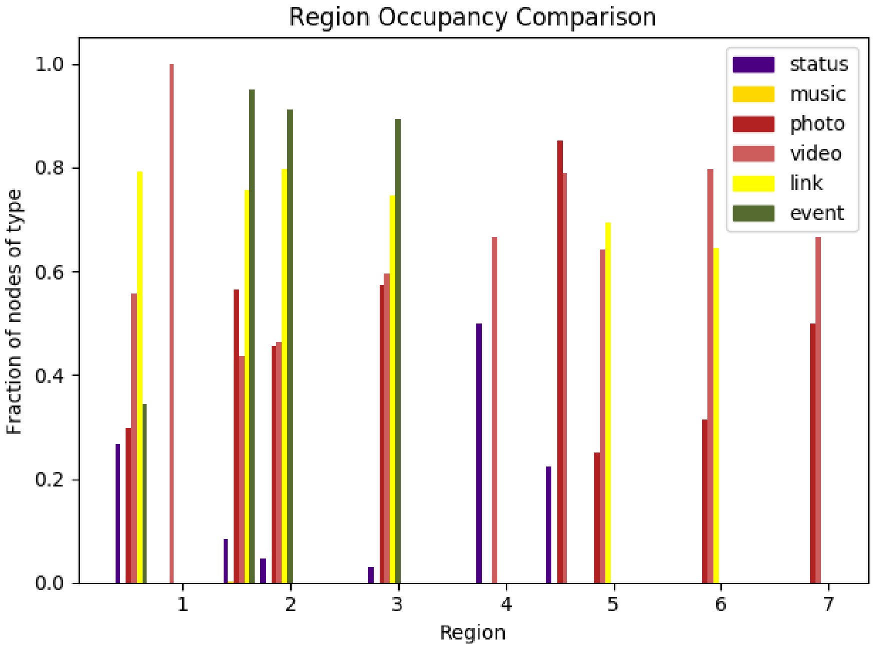

Figure 5: Bar chart comparisons of Region occupancies depending on the type of post (y scale is proportional to actual fraction). Glasgow data on the left and Barcelona data on the right for each region.

One of the most interesting features of this graph is the presence of the 'event' posts. It is clear that in this data set, albeit a small size, all event posts do not appear above Region 3. This could be further analysed by comparing this data set to a randomly generated one and finding the probability and hence significance of this arrangement. It is also interesting to note that Glasgow and Barcelona have an almost identical distribution in Region 2. Glasgow also has a much higher spread of types of post present in Region 1 (due to the fact it has many more posts in this region).

An important comparison to make is between the data obtained through this analysis with the actual projected network.

It is clear that the notion that Glasgow is a more poorly interconnected network agrees with the networks in *Fig.2*. This is an important result as it shows that the analysis performed agrees with intuition and basic visual analysis.

Conclusion

Both visual and statistical analysis of data from Glasgow and Barcelona appeared to agree with the notion that Barcelona has a much more interconnected network of music fans. This project only looked at a small fraction of music fans (around 100,000 in each city), and thus cannot guarantee a proper representation of the full population. However, this project has produced some interesting results which I would like to explore further.

Further Work

There are several aspects of this project which, given more time, I would like to pursue further. These include an exploratory analysis of more major cities across the World to try and establish a pattern and mechanism for classifying cities in terms of their music fans. This could perhaps prove useful as a way of better organising festivals and music concerts in various locations, by driving more traffic on social media with targeted advertising: allowing organisers to use the city's network of music fans to their advantage. It may also be interesting to compare the community of music fans to the community of another set of 'artistic' fans in the same city: this may allow a determination to be made about whether this community structure is specific to the type of fan or to the geographic location of the fan.

Extra Work

Paris

Music community analysis as performed above for Glasgow and Barcelona was briefly performed for Paris. It can be seen in *Fig.6* that the data for Paris appears to follow very closely that of Glasgow, with one significant difference. Paris has a reasonably sized proportion within Region 3, whereas Glasgow has none.

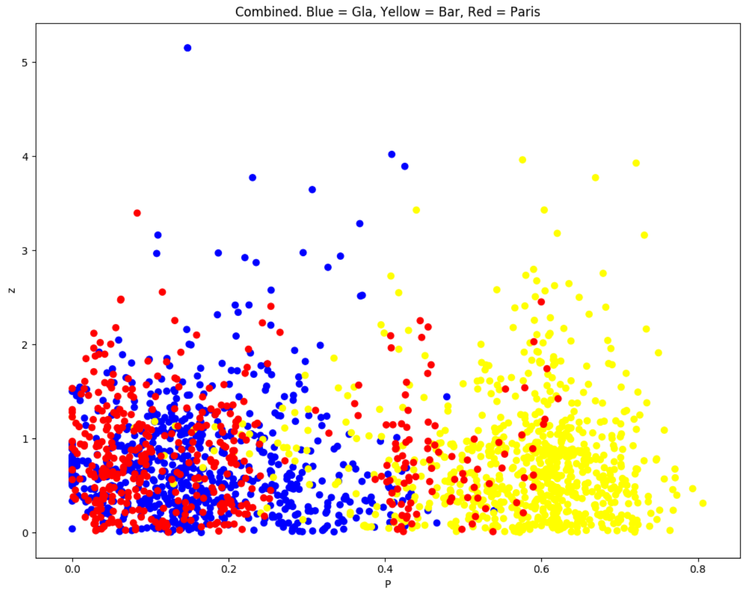

Figure 6: z-P plot for Glasgow, Barcelona and Paris combined (blue data for Glasgow, red for Paris, and yellow for Barcelona).

It is also interesting to note that similar analysis on the type of post, as performed above, yields the same result regarding 'event' posts. Paris also does not have any 'event' type posts beyond Region 3.

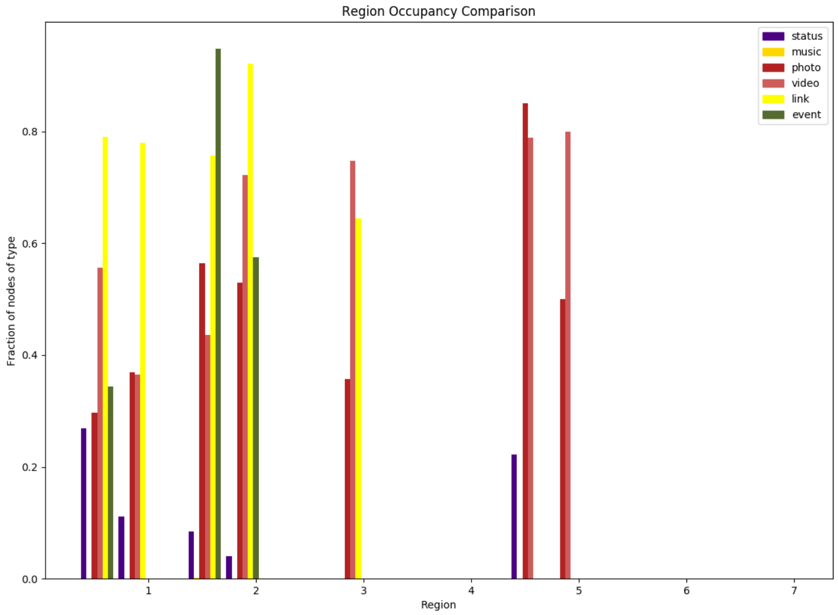

Figure 7: Bar chart comparisons of Region occupancies depending on the type of post (y scale is proportional to actual fraction). Glasgow data on the left and Paris data on the right for each region.

University Facebook Page Comparison

Brief work was also done on the social network surrounding various University Facebook Pages. For comparison, data was collected over 18month periods for Imperial College London, Universitat de Barcelona and the Massachusetts Institute of Technology. To better understand and try to classify the community around these pages, the user engagement per post was analysed. This involved basic counting statistics based on when a user 'liked' or 'commented' on a post: referred to from now on as 'user engagement'. Histograms of user engagement per post were plotted, in order to try and find a mathematical description of the data. Visually the data suggested a power law fit may be appropriate (see *Fig.8*).

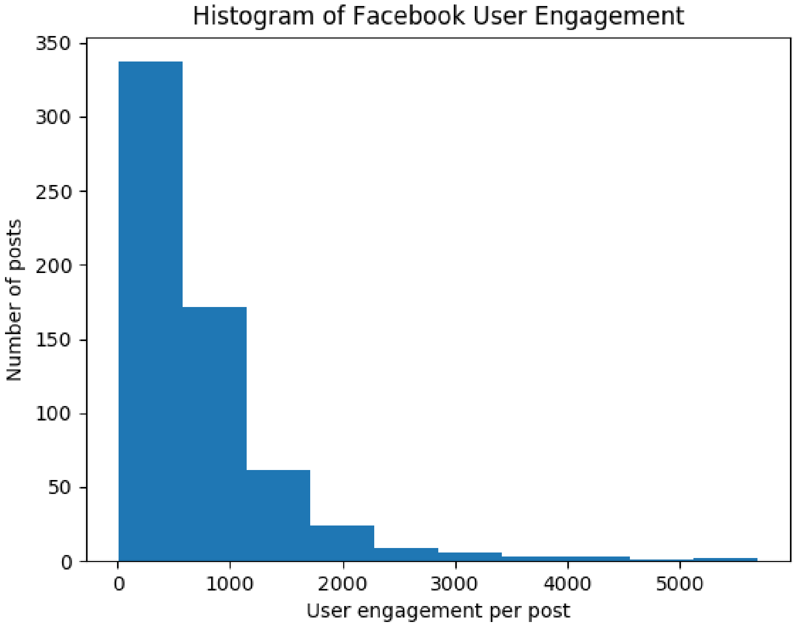

Figure 8: User engagement per post histogram, example.

Upon deeper analysis of these histograms, it was found that the best fit in all cases was a lognormal (see *Figs. 9,10 and 11* for examples for each University). These figures shows that a lognormal fit works best with the empirical data in all cases (when compared to an exponential, stretched exponential and power law).

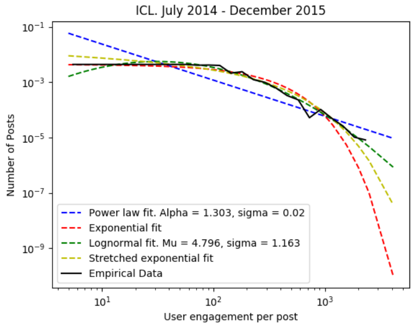

Figure 9: Imperial College London User Engagement Data

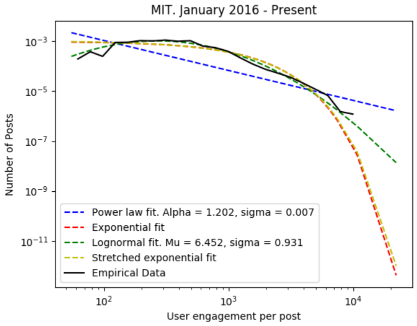

Figure 10: MIT User Engagement Data

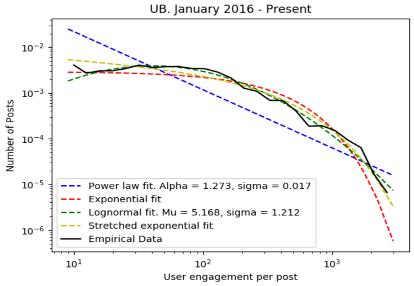

Figure 11: Universitat de Barcelona User Engagement Data

It would perhaps be interesting to look deeper into this interesting result. Lognormal fits do not appear too often, and perhaps these fits could be a way of better identifying the 'best' posts by a University in order to maximise their audience. It is also interesting to note that in a (very brief) comparison with the data above for music pages, it was found that music pages tend to follow either stretched exponential or lognormal fits. I would like to explore further which bands' data lends itself to each type of fit, and why.

References

[1] Spotify Insights Website: https://insights.spotify.com/us/
[2] Newman M.E.J. 2001. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E, Vol 64. 016132
[3] Guimera R, Sales-Pardo M, Amaral L.A.N. 2007. Classes of complex networks defined by role-to-role connectivity profiles. Nat Phys. 2007 ; 3(1): 63-69.
[4] Guimera R, Amaral L.A.N. 2005. Cartography of complex networks: modules and universal roles. Journal of Statistical Mechanics.

I would like to thank everyone in the Complexity Lab Barcelona, Universitat de Barcelona for being so welcoming and helpful during my time here. Specifically I'd like to thank Dr's Luce Prignano and Emanuele Cozzo for their help and guidance throughout my project. Thank you also to Professor Diaz-Guilera for taking me on this IAESTE placement and allowing me to spend time with your group.