Emerging Technologies
Part 3: Modelling Networks
An analysis of the appearance of characters in The Simpsons

Iain Johnston (C1312579)

Contents

1

Dataset

The dataset that I have chosen to analyse involves The Simpsons, the long running American cartoon television show.

Table 1: Basic statistics about dataset

No. of episodes	546
No. of characters	612
Season covered	1 - halfway through 25

The original dataset comes from `https://github.com/sghall/simpsons-episode-data` with thanks to Delimited Technologies (`http://www.delimited.io/`), where it is formatted as an sql document. I have adapted this data into two csv files for use with Gephi.

1.1 Edges

The **edges** of the network correspond to an episode in which the two characters, who are linked by that edge, play a main role. This means that even though a character may *appear* in an episode it does not mean that they are involved in that episode according to the dataset. Hence why Homer Simpson is not involved in every episode in the data, even though it is very likely that he appears, at least briefly, in almost every episode.

I have chosen to use these edges as I believe it will allow a broad overview of the universe and the characters within it. Furthermore one episode would not give a large enough dataset for any real conclusions to be drawn.

To adapt the edges I wrote a python script to transform the original data into a more suitable set for visualisation. This code can be found in *data.py*. This is because the original dataset (an SQL table called episode_character) contained a list of tuples in the form (episode_id, character_id). An extract of this original file is shown below (the full file is attached):

simpsons_ep-char.csv

```
episode_id,character_id
1,1
1,2
1,3
1,4
1,32
1,33
1,34
...
```

After running the Python script the data was transformed into a list of character id's where a row corresponds to the fact that the two characters (whose id's are present) appear in an episode together. Hence this list is a list of **edges** between two characters (**nodes**) in the

network. Finally as this list contains duplicate rows (as two characters can appear together multiple times), it was necessary to parse the file and transform the duplicates into a single entry with a weight corresponding to the number of duplicates of this entry. This was done by using another python script (*data2.py*). Both python scripts are attached.

An extract of this transformed dataset is shown below (the full file is attached):

simpsonsEdges.csv

```
Source,Target,Type,Weight
2,1,Undirected,441
3,1,Undirected,438
3,2,Undirected,393
4,1,Undirected,402
4,2,Undirected,356
4,3,Undirected,380
5,1,Undirected,112
...
```

1.2 Nodes

The **nodes** of the network correspond to characters in the Simpsons, this means main characters such as Homer Simpson but also celebrities playing themselves in a single episode. To view the full list of characters (**nodes**) the attached file *simpsonsNodes.csv* contains character names and a description of them (which is unused in this analysis).

The choice for characters representing nodes was quite simple. I believe that the most interesting conclusions that could be drawn from this dataset are about the characters themselves. It would be possible to swap the nodes and edges and represent episodes as nodes and this would give a different set of results and a different set on conclusions could be drawn, however I think characters would make a more for interesting investigation.

The file for the nodes, was easier to adapt as it was already in the correct order. For this analysis I did not use the *char_desc* column. The main and important adaptation I needed to do was change the *char_id* column name to *Id*. This is because the dataset I have uses 1-based indexing and Gephi was giving each row a new 0-based index called Id and hence the edges were missing some links. I re-named the column to Id so Gephi would pick it up. In this file *Id* corresponds to the same values in the Source and Target columns of the edges file. An extract from this file is shown below (full file attached):

simpsonsNodes.csv

```
Id, char_name, char_desc
1,'Homer Simpson','Homer Jay Simpson is the patriarch of the eponymous family.
He is voiced by Dan Castellaneta. As the family\'s provider, he works at the
Springfield Nuclear Power Plant. Homer embodies several American working class
stereotypes: he is crude, bald, overweight, incompetent, clumsy, lazy, a heavy
drinker, and ignorant; however, he is essentially a decent man and fiercely
devoted to his family. ','http://en.wikipedia.org/wiki/Homer_Simpson'),
2,'Marge Simpson','Marjorie \"Marge\" Simpson is is the well-meaning and extremely
patient matriarch of the Simpson family. She is voiced by Julie Kavner. With her
husband Homer, she has three children: Bart, Lisa, and Maggie. Marge is the
```

```
moralistic force in her family and often provides a grounding voice in the midst
of her family\'s antics by trying to maintain order in the Simpson household.',
'http://en.wikipedia.org/wiki/Marge_Simpson'),
```

After loading the dataset into Gephi it was obvious that there were some duplicate nodes; most likely caused by accidental duplication in the original dataset. However that was easily rectified by merging those nodes in the Data Laboratory. Furthermore for this analysis I decided to remove all nodes with a degree of 0 as these correspond to characters who have no connections in the dataset so are not useful to analyse but may cause some statistics to be skewed. This was done in Gephi so the original dataset still contains those nodes.

2

Analysis

Some basic information for the network is shown in the table below:

Table 2: Basic statistics

Nodes	612
Edges	6665
Max node degree	592
Max edge weight	441

Firstly this network is an undirected graph, which means that a link between two nodes is just a connection between those characters. This is because I could not see a correct and useful way of classifying edges into directions as the concept of the direction of a connection doesn't make sense in this dataset.

Two statistics of the network that may be interesting to start by looking at are **node degree** and **edge weight** as they may be able to help reveal interesting parts of the network by filtering and analysing their distributions, without having to contend with the full network.

2.1 Node degree

The degree of a vertex of a graph is the number of edges incident to the vertex[1]. In this network, for a given vertex (= *node* = *character*), the degree of that node is the number of other characters with which they appear in an episode.

In terms of The Simpsons, this idea can be expressed as a sort of popularity rating of any given character, as that character appears with others very often.

The degree distribution can be taken further by taking into account the weight of the links at each node. This becomes the weighted degree distribution shown in figure 1

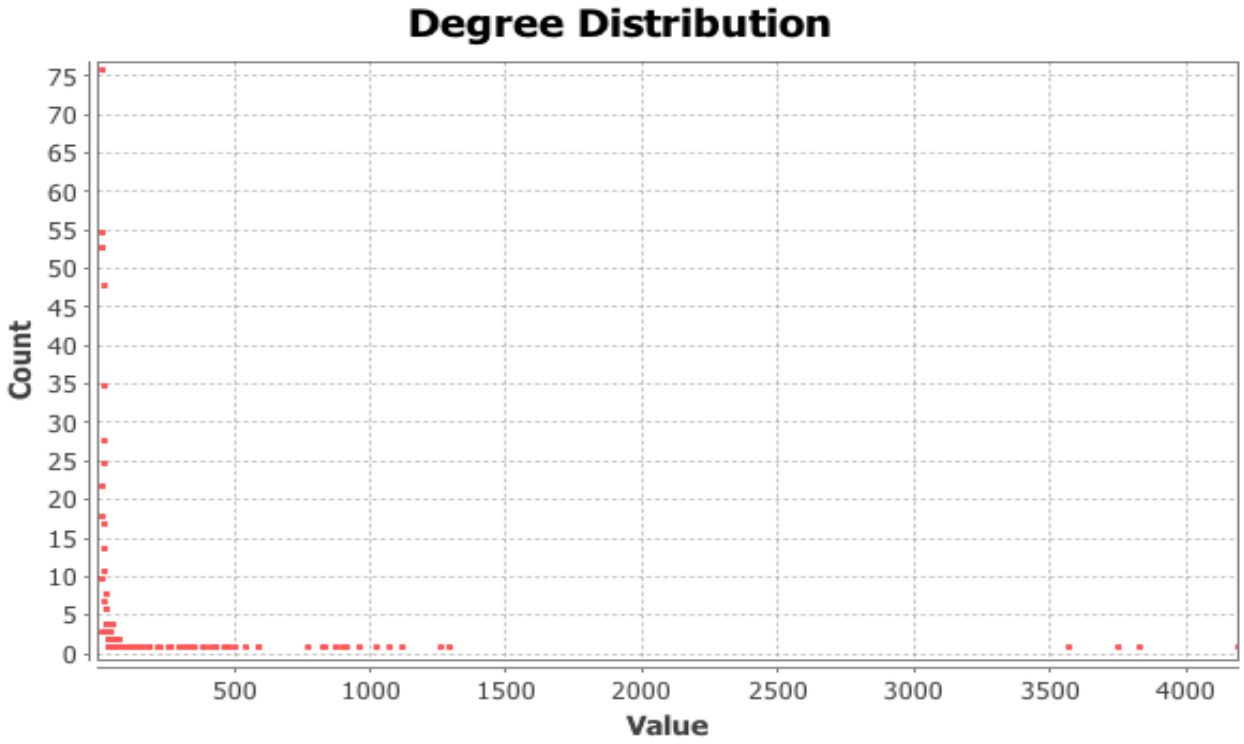

Figure 1: Weighted Degree distribution of the network

We can see that there are few nodes with a high degree and many with a very low degree. This suggests that there are only a few very popular characters in The Simpsons and many characters who don't appear with others very often. Furthermore as this is the *weighted* degree distribution over the network, it can also be said that there are few characters who have lots of relationships with other characters, which is what the edge weight reveals. The edge weights and what they mean is explained in the next section of the analysis, as they are interesting in their own right.

Another way of showing this is by applying the degree distribution to the node size as in figure 2. This helps to take the information on the chart and apply it more obviously to the dataset. The network is shown below with the distribution applied to the size of the node, and then laid out. From this it can be seen that the majority of nodes are not exceptionally popular, and only a few are very popular.

N.B *popular* in this case really refers to the mixture of the node degree and edge weight into a weighted degree, which in turn leads to the idea of those measures in the Simpsons universe being the "popularity" of a character.

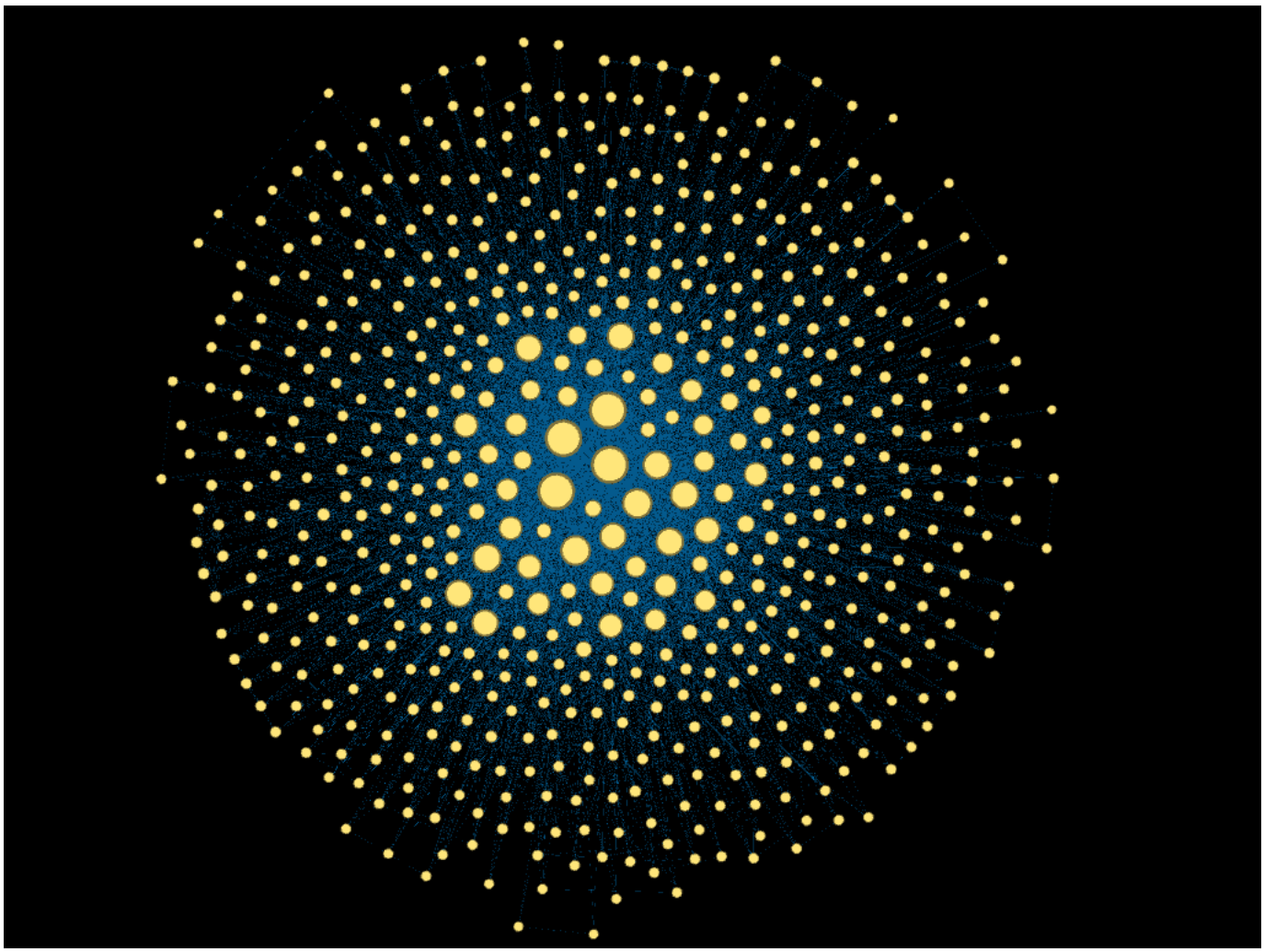

Figure 2: Weighted Degree distribution of network (size) with edges

The chart in figure 1 can also help to filter the network. We know that the maximum degree is 592 (*Homer Simpson*), and from the chart we can see that a possibly interesting point at which to remove a large part of the network that is uninteresting, is at around 100. Figure 3 is the network, filtered for degree greater than 100.

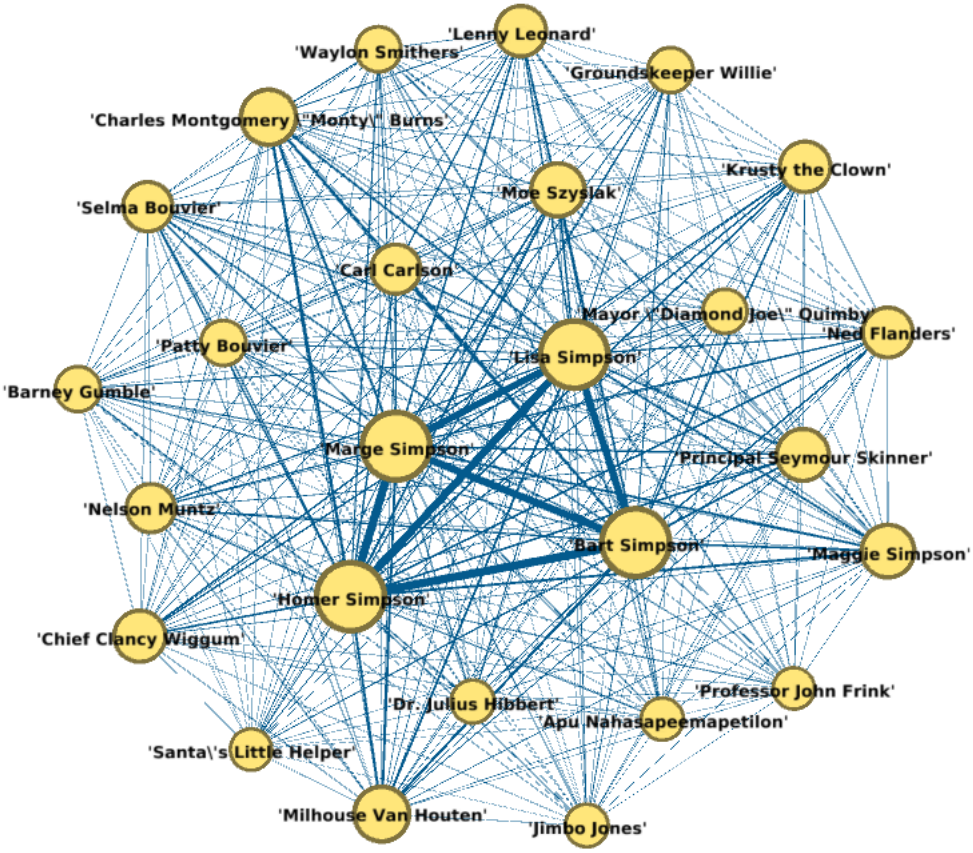

Figure 3: Nodes with a degree of 100 of greater

From this we can see the most popular characters, and all of these characters are not surprising to anyone who has watched The Simpsons. Furthermore from figure 3 it can be seen that these characters are all quite well connected, which again is not surprising as they are all very important characters in the universe.

There are a few small surprises revealed by this filter, namely that Groundskeeper Willie has quite a high degree. This is a surprise because my first thought is that he is the type of character who usually has a small sphere of influence with other characters, and it is hard to pin-point many episodes in which he plays a main role. However the network suggests he is more important than I first thought.

Another interesting conclusion that can be made is that after the 4 Simpsons (Homer, Marge Bart and Lisa), the next 4 characters with respect to degree, are *Milhouse*, *Mr. Burns*, *Krusty* and *Principal Skinner*. This is not surprising as they are commonly cited as other main characters. This is still a nice result as it allows us to be able to see who the most popular characters are, other than the Simpsons themselves.

2.2 Edge weight

The edge weight is a value given to an edge between two vertices. In this network, the weight on an edge is the number of episodes in which the source and target nodes appear together.

In terms of The Simpsons, this can be thought of as the *relationship* between two characters as a higher weight between them means they appear together often. Of course they can appear in an episode together but never actually interact which is why it is necessary to understand the underlying dataset, however at the higher end of weights it is likely that the relationship is valuable and not an accident.

The first thing we can do with regards to edge weight is remove the main 4 characters as they have the highest edge weights so it is less interesting to look at them. If we filter the network using the query and settings shown in figures 4, 5, 6, we can see the next character pairs based on their edge weight.

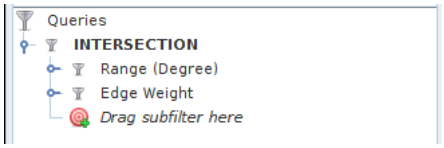

Figure 4: Filter query

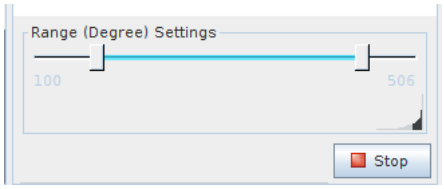

Figure 5: Degree range

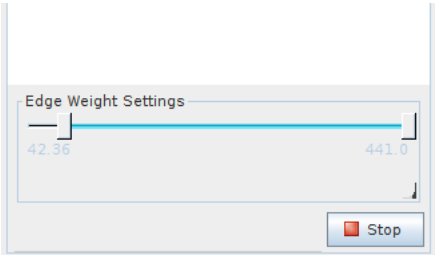

Figure 6: Edge weight range

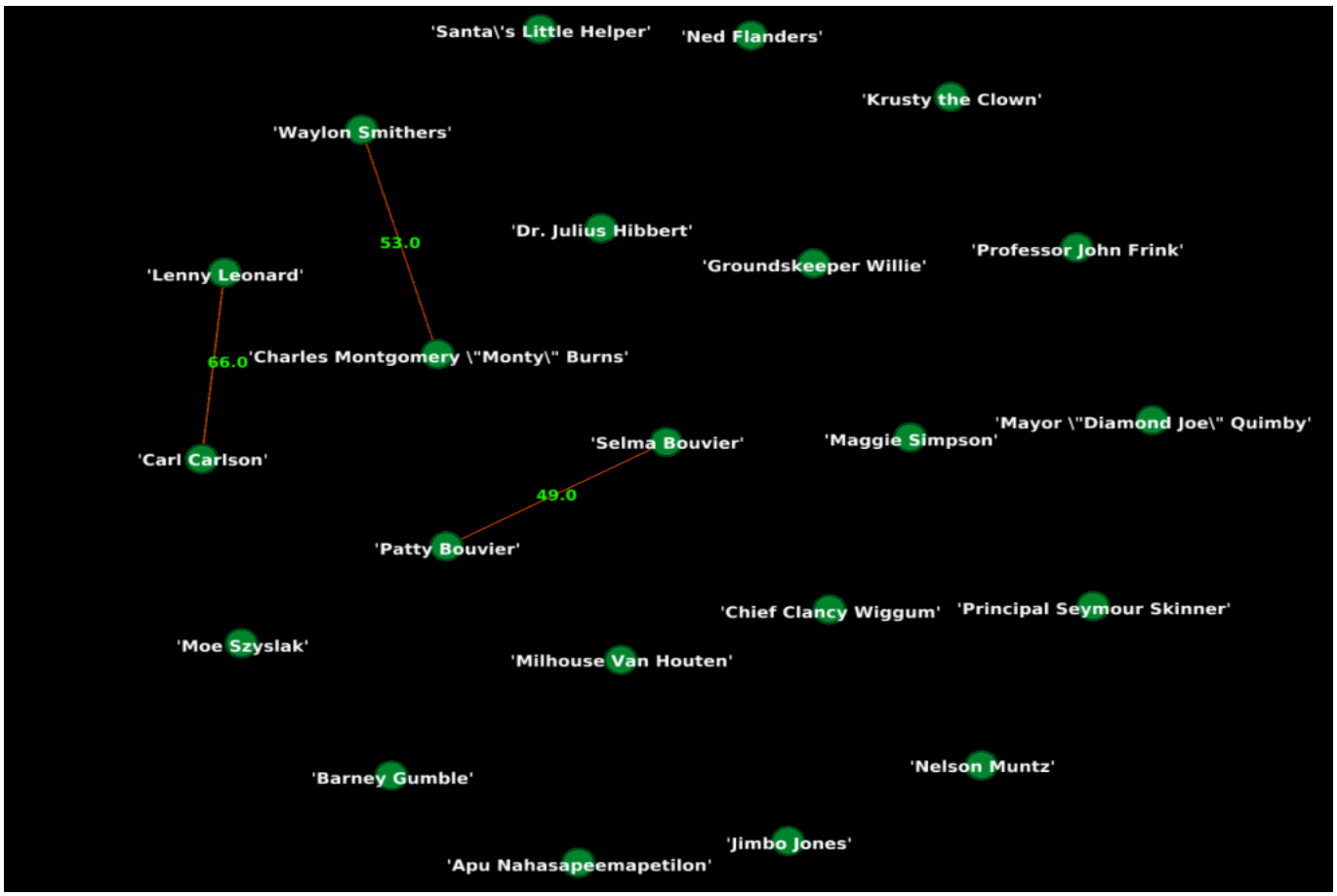

Figure 7: Network with intersection query filter applied

As the weights correspond to the number of episodes in which they appear together, after removing the main family characters from the dataset, we should be able to find interesting pairs. This is both because the main family is very well connected to each other, but also because they are also very well connected to the other characters, especially the other non-family main characters.

From it we can see that the characters who appear the most together are *Lenny and Carl* (66 episodes), *Smithers and Burns* (53 episodes) and *Patty and Selma* (49 episodes).

Lenny and Carl are two of Homer Simpsons friends and usually appear together, and usually also with other characters such as Moe and Barney. The relationships between Lenny and Carl is quite well known among fans, and has been referenced on several occasions, especially in more recent episodes; usually as part of a joke that they are never seen apart. This analysis backs up this widely help joke.

Figure 8: Lenny (L) and Carl (R)

What is even more interesting is the "strength" of the relationships. In a dataset that contains 546 total episodes, for *Lenny and Carl* to appear together in 66 episodes means they appear together in 12% of

episodes. Comparing that to how many times they appear in episodes with the main 4 Simpsons as shown in table 3, *Lenny and Carl* seem to have a very strong relationship.

Table 3: Lenny Leonard and Carl Carlson - and The Simpsons

Character 1	Character 2	No. of episodes together
Lenny	Carl	66
Lenny	Homer	85
Lenny	Marge	73
Lenny	Bart	68
Lenny	Lisa	69
Carl	Homer	80
Carl	Marge	72
Carl	Bart	65
Carl	Lisa	63

Figure 9: Smithers (L) and Burns (R)

The relationship between Waylon Smithers and Mr. Burns is another one that tracks in the universe. Smithers is Mr. Burns assistant and they are rarely apart. This pair differs from Lenny and Carl in that Mr. Burns is actually a very important character in his own right and appears as the main antagonist in episodes quite often. Smithers on the other hand is alone in episodes very rarely and almost never plays a leading role.

As with Lenny and Carl, jokes have been made in many episodes about how Smithers feels towards Mr. Burns and their dependence on each other has even been the main plot point of an episode (S7E17 - *"Homer the Smithers"*)[2]

Patty and Selma Bouvier are twins sisters of Marge Simpson and are another pair who almost never appear on their own. One recurring joke that has appeared many times is that they are very difficult to tell apart as their voices and mannerisms are very similar, however they do have distinguishing characteristics such as their dress colour.

In contrast to both Lenny and Carl and Smithers and Burns, they do appear as main protagonists of episodes on a couple of occasions (namely *Selma* in S16E12 - *"Goo Goo Gai Pan"*)[3], but they also appear very often with the main family but almost always together.

Figure 10: Patty (L), Selma (R)

2.3 Number of Triangles

The number of triangles of a node is the number of times they are fully connected with two other characters. This distribution looks interesting as it can help to back up some of the conclusions made about important characters based on their Degree.

This can be seen quite nicely in figure 12, where the network has been coloured according to the number of triangles, using Gephi's node colouring tool (shown in figure 11).

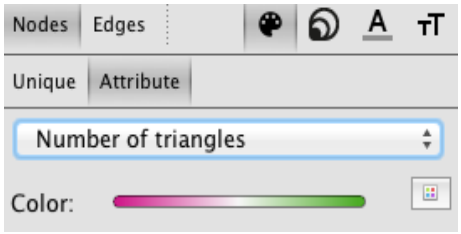

Figure 11: Gephi colouring tab

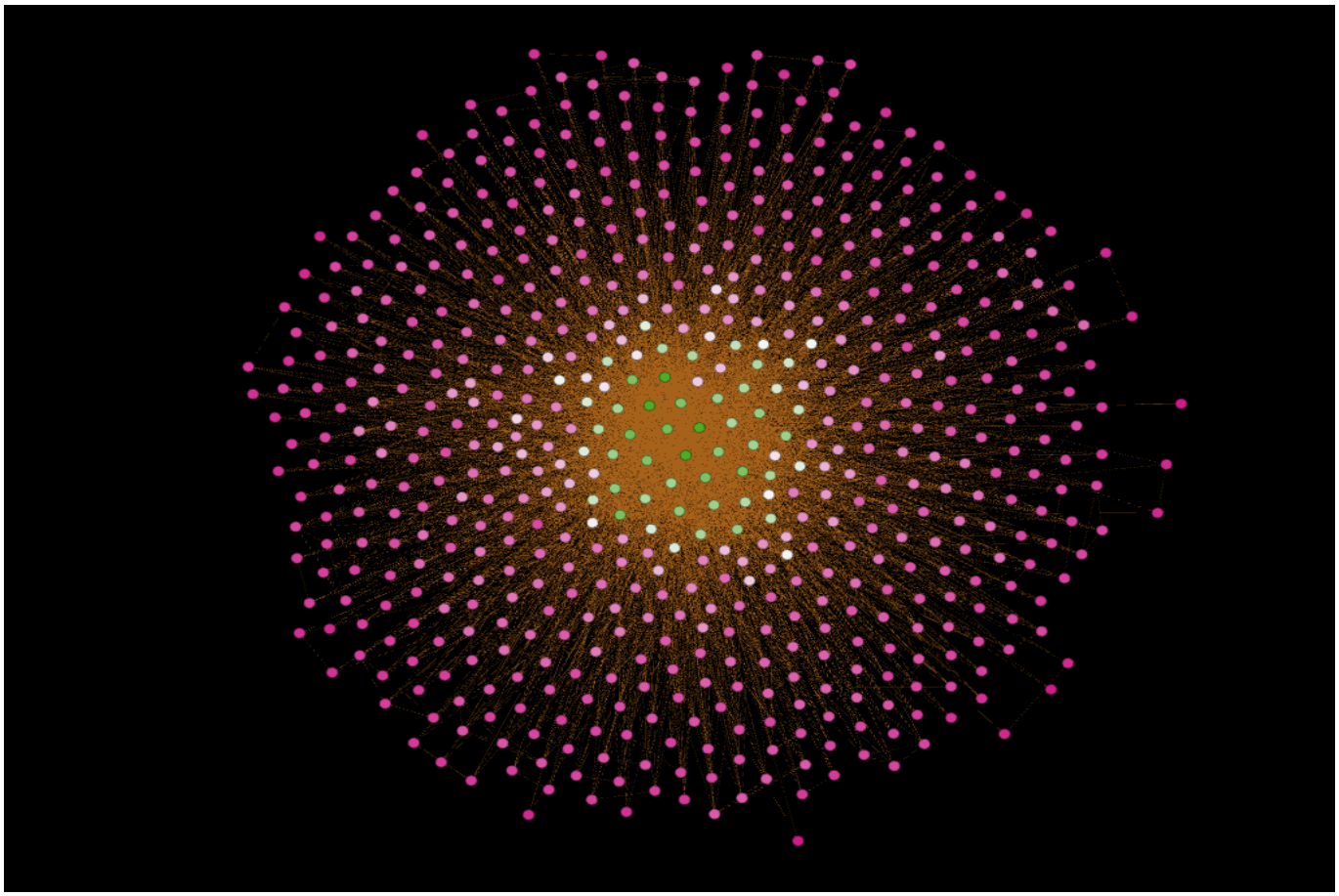

Figure 12: Network with no. of triangles (colour), pink = low, green = high

Figure 12 suggests that only a small number of characters are very well connected. As the dataset contains 612 characters then it is not surprising that most of them are not highly connected. A large portion of the characters only appear a handful of times and will not have appeared with the same characters often enough for them to show as a strong part of the distribution.

Figure 13 shows the same network as figure 12 but it has been filtered to show nodes with number of triangles greater than 1100. Moreover the nodes are labelled with the number of triangles.

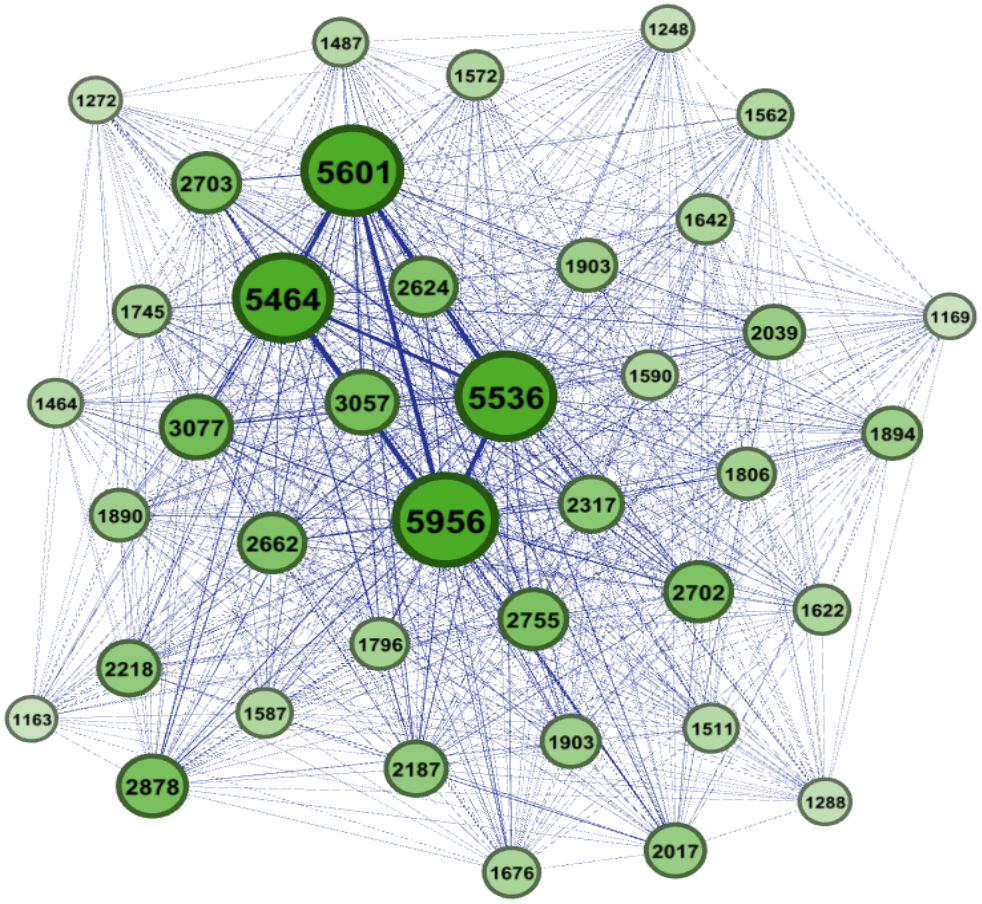

Figure 13: Network filtered for high no. of triangles, with size and colour as no. of triangles

From figure 13 it can be seen that once again the Simpsons family have the highest, and again this is not surprising as they are the most important characters.

What may be more interesting to look at is the next set after the Simpsons have been removed. This is shown in figure 14.

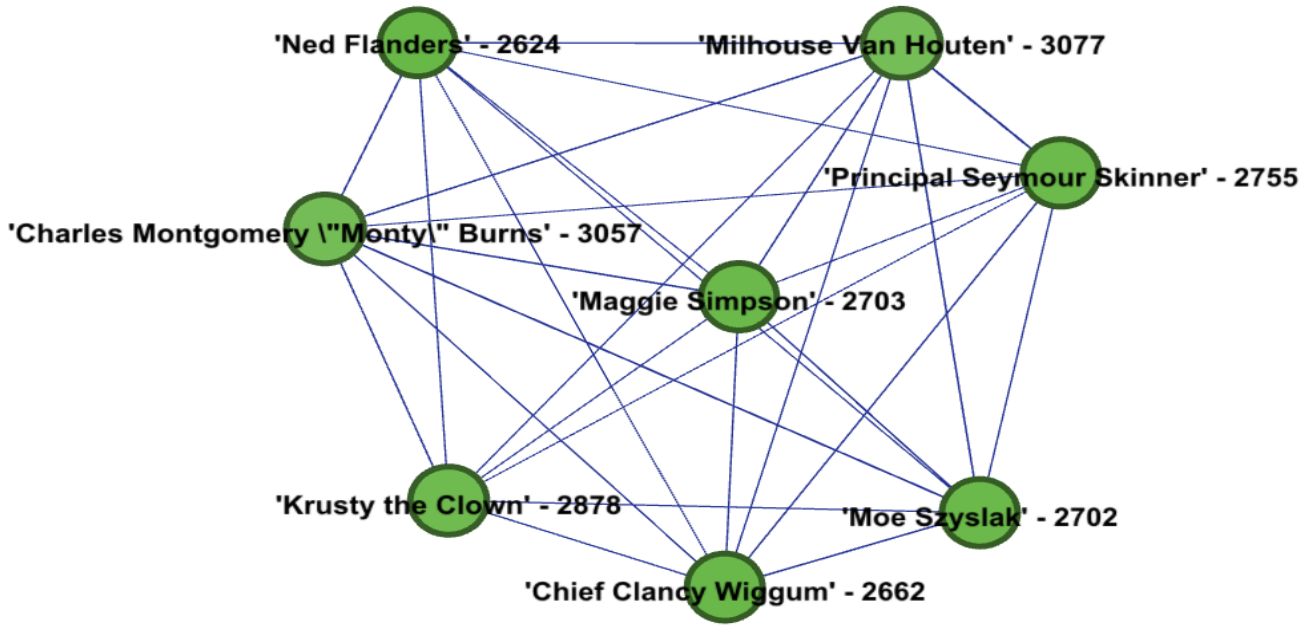

Figure 14: Network filtered for no. of triangles > 2500 and highest 4 nodes filtered out

What is most interesting here is that these characters closely match those that were concluded to be important based on their node degree. So this measure backs up the conclusions made about the dataset in that analysis. Moreover the characters are in a similar distribution as node degree, so it can be shown that those measures are quite closely related. This makes sense as the number of triangles is in many ways an extension of that node's degree.

This also lends credence to the fact that these characters have important connections outside of the main Simpson family, as their values are still very high, meaning they appear with non main characters very often too.

2.4 Modularity and Centrality

The modularity of a network is a measure of how well it can be split into communities. It is generally thought that a high value for the modularity means that the network has a complex underlying community structure. These communities can sometimes have significant meaning in the network.[4][5].

Although this network does not have a high modularity value it can still be interesting to analyse the clusters that can be found. In this case, it is necessary to only analyse the main groups as there most likely isn't a hugely complex underlying structure in this network.

If the modularity is taken with the Betweenness Centrality measure, some communities within the network and their most important nodes can be shown. *Betweenness Centrality* is a measure

of how often a node appears on shortest paths between nodes in the network.[6] .This is shown in figure 15.

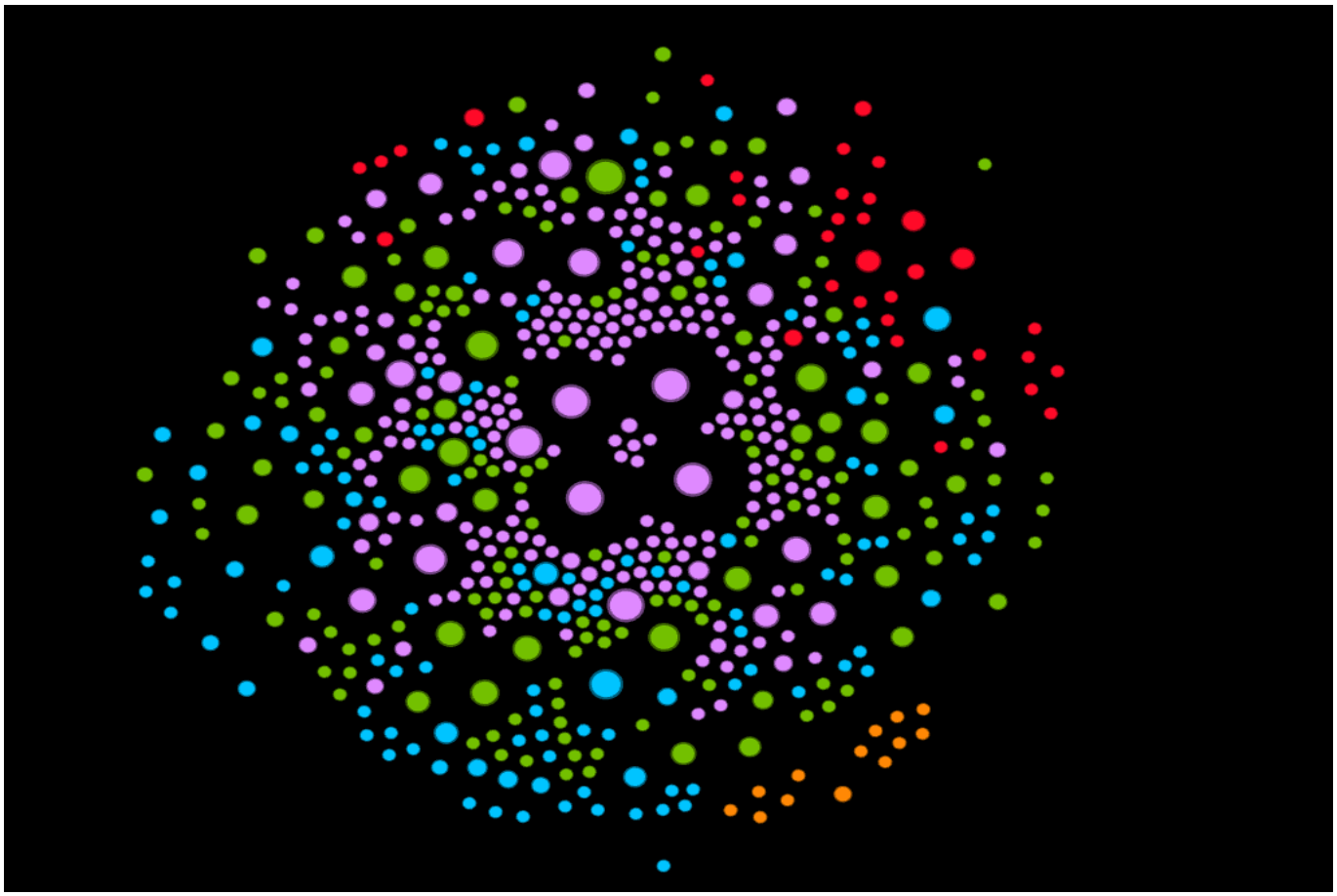

Figure 15: Network with no edges, modularity classes (colours) and Betweenness Centrality (size)

Furthermore the network has been filtered to show only the top 5 modularity classes, i.e. classes with the most nodes. From here we can look at some of the classes to find out more about their members. As mentioned above, we only look at the top 5 classes because of lack of a real complex structure.

In figure 16 we can see modularity class 1, which was coloured pink above. This community is quite large and contains the main 4 Simpsons. This leads to the thought that this class corresponds to many characters who appear with main family as the modularity algorithm looks to find nodes that are more densely connected to each other than to the rest of the network. The Betweenness Centrality (size) of the nodes also supports this fact as they are the most influential nodes in the class and they correspond to the main characters.

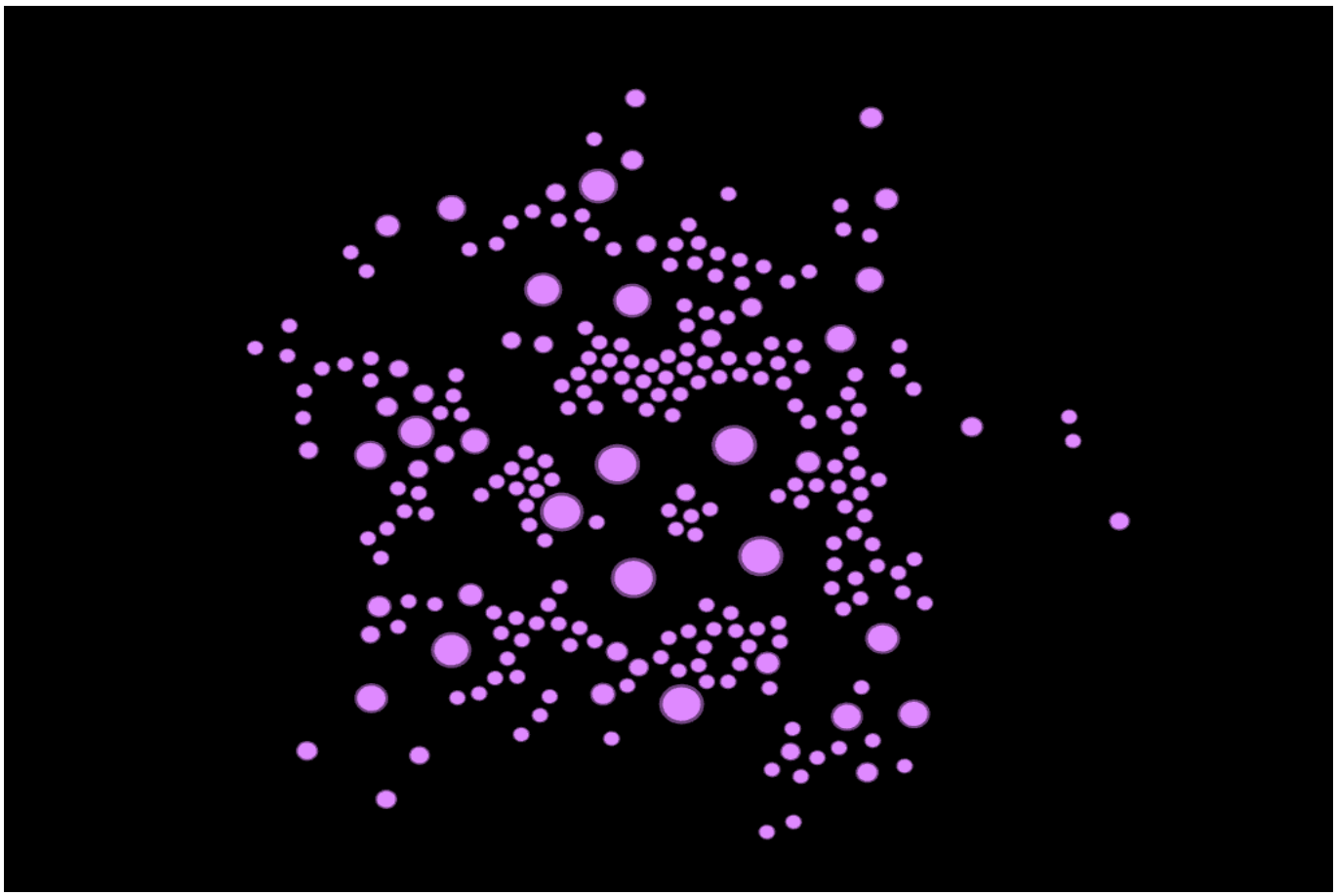

Figure 16: Modularity class 1

This is useful to analyse this group because it can reveal those connections that the other "importance" measures such as edge weight and node degree, may have missed.

The second modularity class (green) is shown below. As with class 1 it is quite large but this time contains only a few influential nodes. These still correspond to regular characters such as *Chief Wiggum* and *Principal Skinner* but not the main family. This leads to the conclusion that this class corresponds to more peripheral characters.

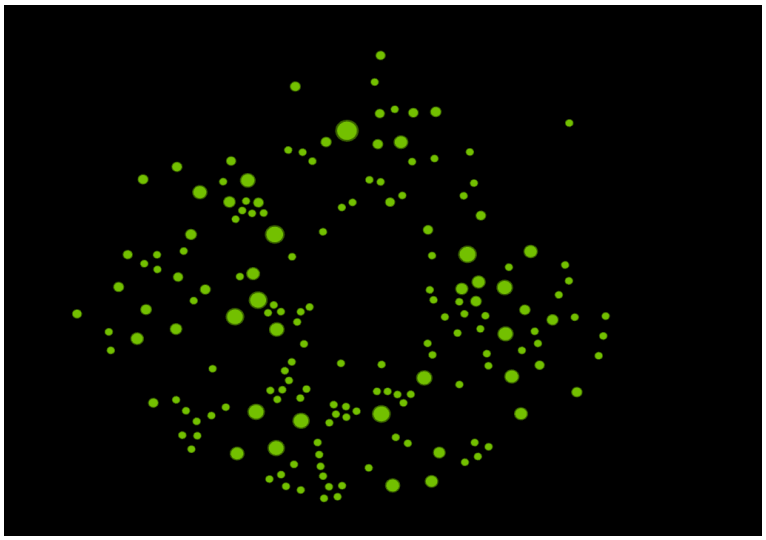

Figure 17: Modularity class 2

Again this view of the network can reveal those nodes who are interesting because they have a stronger connection to the peripheral characters than to the actual main Simpson family.

It is only really useful to look at the first two modularity classes as they make up most of the dataset. Due to the low modularity value, the rest of the classes found will not yield as interesting information as they will be made up of mostly celebrity and one-time characters.

2.5 PageRank

PageRank is an algorithm that measures the importance of each node in the network.[7]. This is another measure that we can use to discover who are the most important characters in the show. The algorithm takes into account both the quantity and quality of the edges between nodes. This means that more important characters are likely to be linked with other important characters more often. This measure is most likely one of the best at finding the important characters as it takes into account parts of the other measures.

The PageRank distribution is shown in figure 18.

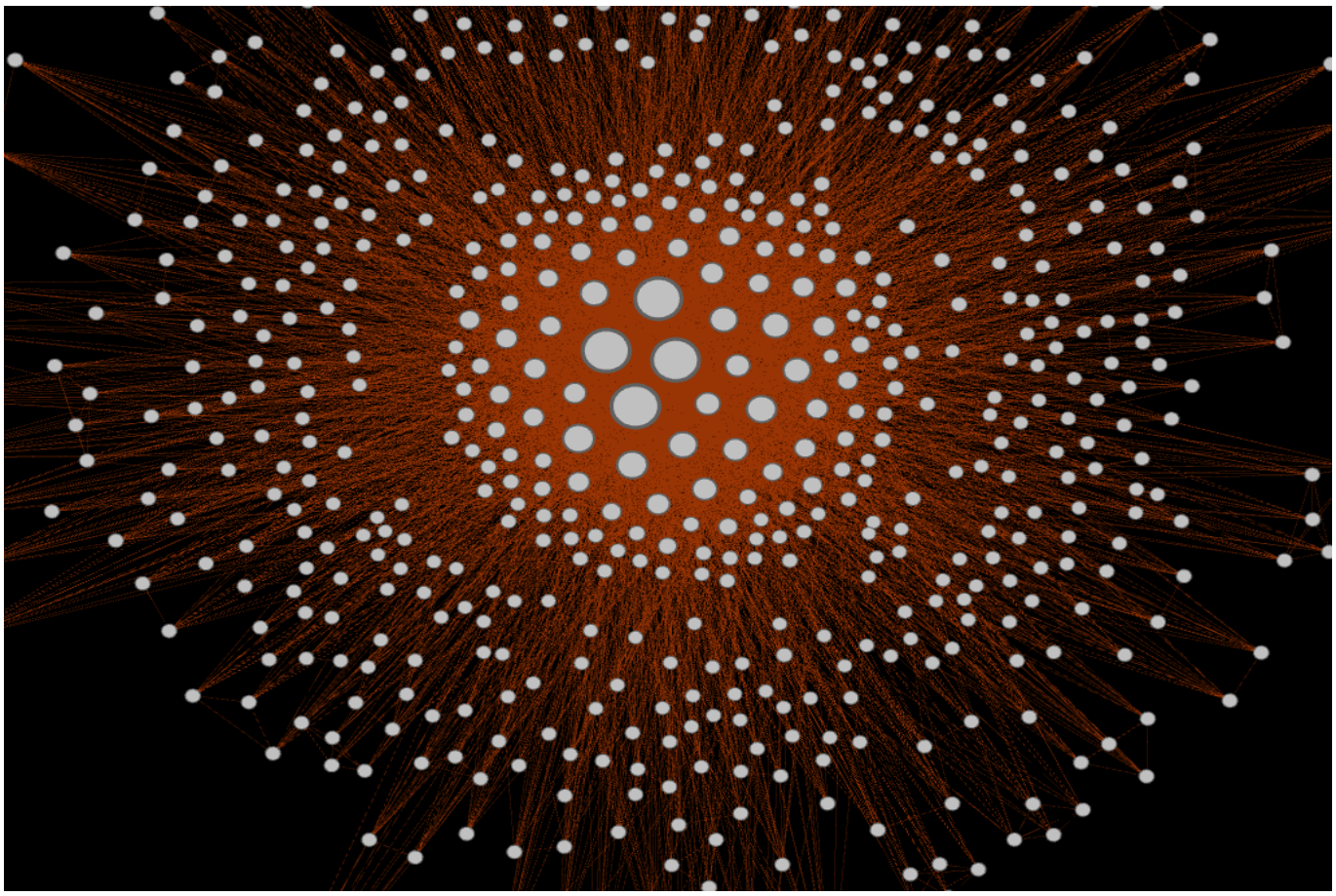

Figure 18: Network with node size = PageRank

This distribution generally seems to follow the other importance measures, in that the main Simpson family have the highest PageRank values, however where it deviates to some extent is with the next set of highest PageRank values.

Whereas node degree and edge weight suggest that the most important characters after the Simpsons are *Milhouse*, *Mr. Burns*, *Krusty* and *Principal Skinner* (in order), PageRank says

that *Mr. Burns* is actually **more** important and that *Principal Skinner* is **less** important than first thought.

Another interesting result is that whereas *Maggie Simpson*, was not rated very highly by the other importance measures, according to PageRank, she is the 8th most important character. This is most likely because PageRank has some understanding of the quality of the edges between characters, where edge weight and node degree do not. This means that *Maggie* is placed higher by PageRank because she appears very often with other important characters (the main family) and not very often without them.

This is an even more interesting result as it is something that node degree and edge weight do not reveal.

Another thing that the PageRank distribution reveals is the fact that outside the small group of very important characters, the rest of the characters seem to sit at about the same level of importance. This is interesting because your initial thought would be that the distribution is more spread out, but figure 18 suggest that this is not the case.

3

Visualisations

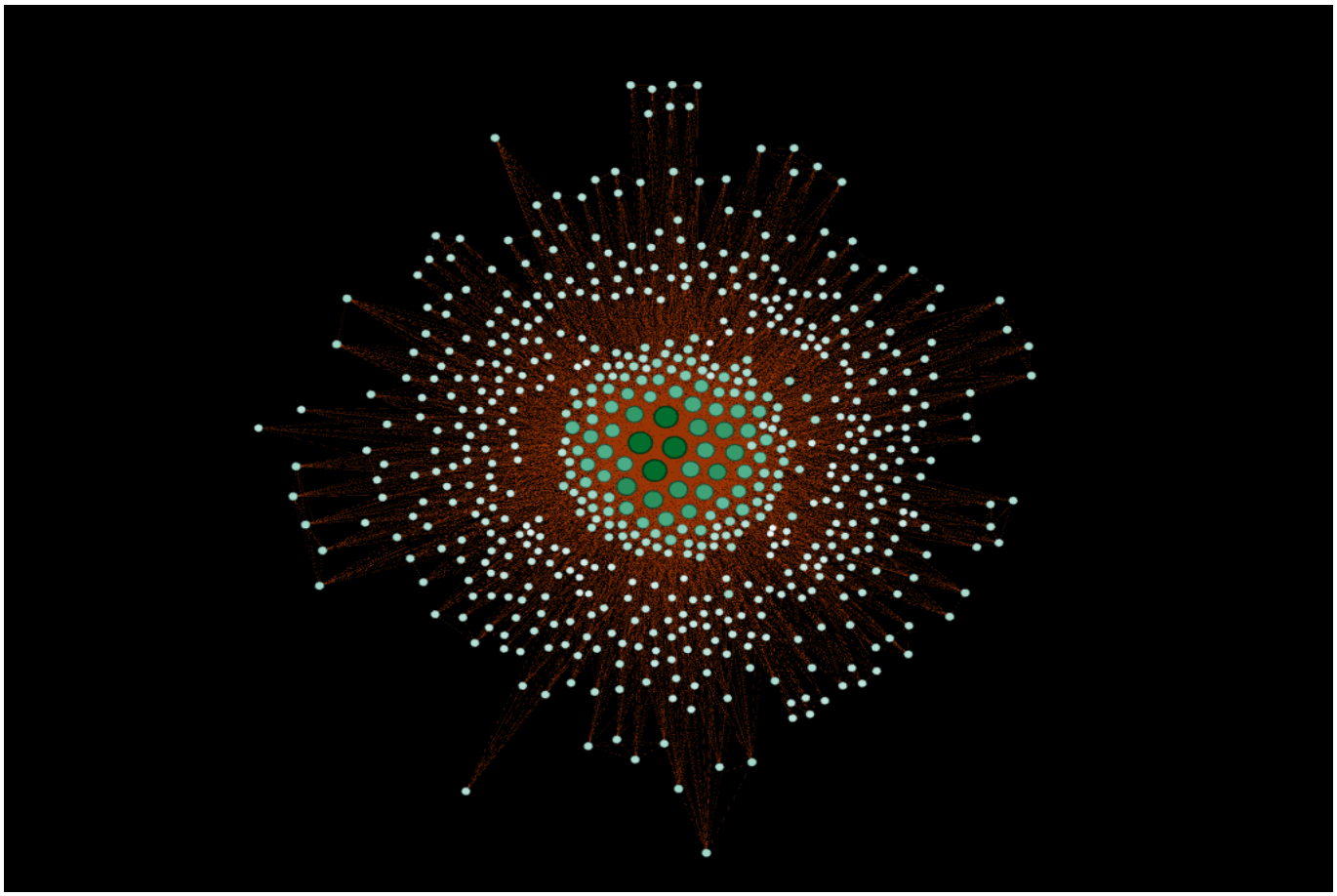

Figure 19: Network laid out with ForceAtlas2/Noverlap. Colour = Weighted Degree, Size = PageRank

This visualisation in figure 19 gives a solid overview of the network and at the same time allows the main properties of this network to stand out. Namely it shows the main family, then the main peripheral characters clustered around them, and the large group of characters who appear little in the show sitting in the fringe.

To achieve this layout firstly the colouring of the nodes is done based on the weighted degree for each node. This allows that the degree of each node, which can help to visualise importance, is clear immediately. The degree and weighted degree of the nodes was discussed in the Node Degree section of the analysis (**2.1**).

Secondly the sizing of the nodes has been done as a distribution of the node's PageRank value. Analysis of the PageRank distribution can be found in the PageRank section of the analysis (**2.5**). It is another measures of the importance of a node and so in this visualisation it helps to add to the power of those nodes that are the most important.

Thirdly edges are coloured depending of the weight of the edge. This has the one drawback in that the most important nodes in the centre actually have many more links than those on the periphery, but it is hard to see that. 5098 (over 75%) of the nodes have a weight of only 1,

and 5628 (around 85%) have a weight of 1 or 2. This means that this visualisation has a very densely packed edge colouring in the middle.

Figure 20 shows the network from a less analytical view. It was produced using the Preview tab in Gephi, but I think that it shows a nice overview of the network. The main property that this visualisation has over figure 19 is that it very obviously shows what dataset this network represents (if the viewer knows the characters). This helps to give a simple and quick entry to the network for any viewer.

The size of the labels correspond to their importance so in some ways it does give a quick overview of the network and the ideas studied in the Analysis section.

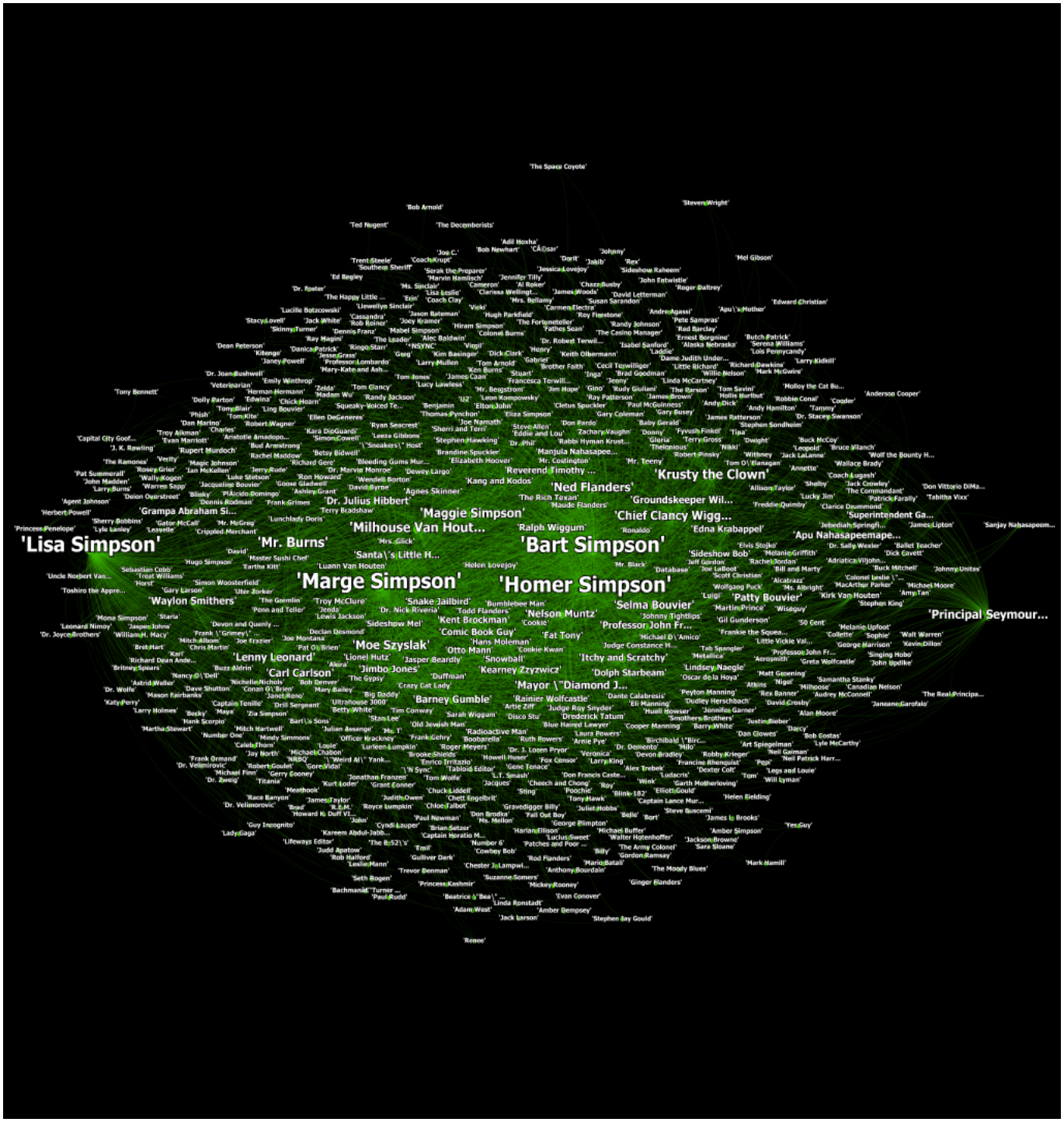

Figure 20: Labelled network using preview tab in Gephi

4

Explanation of Network

My network reveals the connections between the characters in the Simpsons universe, using their appearance together in episodes as a connection. This can be seen as the relative *importance* of each character within the Simpsons universe, as well as helping to reveal relationships between characters. The network has also been able to provide evidence for widely held beliefs about the show.

The most obvious properties that the network tells me is that the main family of the Simpsons (*Homer, Marge, Bart, Lisa*) are very well connected and hence important in the universe. This is not surprising as they are the characters that have been around the longest and are the main thread of most episodes.

The more interesting thing that this network, and the visualisations, reveal is about the peripheral characters and how *important* they are. The network shows that the most important characters after the Simpsons themselves are *Milhouse*, *Mr. Burns*, *Krusty the Clown* and *Principal Skinner*. Even though this might not seem strange to many, it is interesting to see the order that the network suggests for these characters. That is something that would have been hard to take away from the initial dataset.

With reference to the initial dataset, using Gephi to layout and analyse the network has revealed facts that would have been extremely difficult and definitely time-consuming to find out by simply analysing the csv files. Most of the features that the analysis has revealed are significant mainly because they are features of the graph hence they could not have been found otherwise.

All of the findings that have been outlined in the analysis section generally fit the context of The Simpsons universe. For a person who has watched every episode at least once and many episodes multiple times, the network broadly follows what I would have said without looking at any data. Moreover some of the results of the analysis of the network helps to solidify theories about characters that have been held by viewers without actual evidence for many years.

One of the most talked about parts about The Simpsons is the quality of the seasons as time has gone on. Some believe that the best series' were seasons 1- 9 and all those after were hit and miss. This leads to a possible interesting investigation using this dataset. Currently I have been looking at all seasons up to around halfway through season 25.

It would be very possible to reduce the dataset to only those seasons of interest, if you know the episode number of the last episode you want to look at (simply to find on wikipedia). Then you would be able to remove all *episode_id's* larger than that and then analyse the show in the same way as has been done here and compare the importance of characters.

This could lead to conclusions about the relationship between perceived quality of episodes and characters who appear. Moreover one main reason cited for the decline in quality in the later seasons, is that the writers relied too heavily on celebrity guest stars. This could be investigated with this network, by analysing episodes as nodes instead of characters.

5

References

[1] Wikipedia, "Degree (graph theory) — wikipedia, the free encyclopedia," 2016. [Online: https://en.wikipedia.org/w/index.php?title=Degree_(graph_theory)&oldid=751262423; accessed 7-March-2017].

[2] Wikipedia, "Homer the smithers — wikipedia, the free encyclopedia," 2017. [Online: https://en.wikipedia.org/w/index.php?title=Homer_the_Smithers&oldid=765691641; accessed 17-March-2017].

[3] Wikipedia, "Goo goo gai pan — wikipedia, the free encyclopedia," 2017. [Online: https://en.wikipedia.org/w/index.php?title=Goo_Goo_Gai_Pan&oldid=760857049; accessed 17-March-2017].

[4] S. Heymann, "Modularity," 2015. [Online: https://github.com/gephi/gephi/wiki/Modularity; accessed 10-March-2017].

[5] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[6] S. Heymann, "Betweenness centrality," 2015. [Online: https://github.com/gephi/gephi/wiki/Betweenness-Centrality; accessed 10-March-2017].

[7] S. Heymann, "Pagerank," 2015. [Online: https://github.com/gephi/gephi/wiki/PageRank; accessed 10-March-2017].