Network Algorithms and Narratives

R. Walport

Social network analysis sits somewhere between graph theory and sociology. In this type of modelling, people, organizations or whatever entity you choose are represented as nodes and the relationships between them as edges. Throughout this study I will consistently use the terms node and edge for these graph elements (there are several synonymous words for both, such as vertex and line). How we define the nodes and what constitutes a relationship are ambiguous questions and how we tackle them radically affects our results.

The first question before we even begin our analysis is why might network graphs be useful? Ulrike Gretzel, professor of Marketing at the University of Wollongong puts it thus, “Social network analysis is based on an assumption of the importance of relationships among interacting units … Actors and their actions are viewed as interdependent rather than independent, autonomous units.”[1] This essentially means that by making simplifying assumptions, the theory goes that we can derive real knowledge from modelling the world through graphs.

The rise in popularity of social network analysis rests heavily on recent online social networking platforms. These provide ready made networks with clearly defined nodes and edges, inherent in the structured nature of the platforms (e.g. On Facebook you have friends, restricted friends etc.). But the ideas and potential of these techniques moves beyond just these online platforms. The first clear work was done long before Facebook, in the 30s[2], and some of the developments have been used retroactively such as in the case of the paper “Social Movements and Network Analysis: A Case Study of Nineteenth Century Women’s Reform in New York State”[3].

The academic study of graphs is relatively sophisticated and developed. The algorithmic work is sound, using the correct formula you will get a minimum spanning tree for example. The question of what you actually get out of these analyses is more difficult. Can we understand our world better through them? Are there stories buried in these graphs that cannot be immediately seen?

In this study I will investigate what algorithms can do for narrative discovery. Looking at various algorithms in action on a single data set, exploring what can be learnt, what makes sense and what’s practical.

[1] http://lrs.ed.uiuc.edu/tse-portal/analysis/social-network-analysis/

[2] The Development of Social Network Analysis: A Study in the Sociology of Science, Linton C. Freeman

[3] Social Movements and Network Analysis: A Case Study of Nineteenth-Century Women's Reform in New York State, Naomi Rosenthal, Meryl Fingrutd, Michele Ethier, Roberta Karant and David McDonald, American Journal of Sociology, Vol. 90, No. 5 (Mar., 1985), pp. 1022-1054

For this study I have selected a graph of around 75 nodes and 200 edges. Much smaller than many potentially interesting graphs (a social web graph could be on the order of millions), the benefits for our purposes are that we can draw a useful graph without pruning any nodes (i.e. we get the whole data set visually). This is something that often isn’t true in practice where journalists have to make decisions about what to show. This pruning step is clearly crucial for creating a narrative but is a tangential step for us in terms of applying algorithms to graphs to find knowledge.

We will explore a graph based on the novel of Les Miserables by Victor Hugo created by D. E . Knuth[4] (the raw data is attached to this article)! This graph is generated by looking at co-appearances of characters within single chapters, that is: incidences of characters appearing in chapters together. Les Mis is particularly useful in this respect because of Hugo’s structure, breaking the book up into more than 360 short chapters. As a result any coappearance fairly guarantees that the two characters meet (unlike most novels with longer chapters, where multiple scenes might take place in a single chapter). The hope is that by looking at the graph and applying algorithms, we can detect important players, their relationships and maybe even their influence, without using knowledge of the novel (which we can then confirm by applying our knowledge in the manner of a training set).

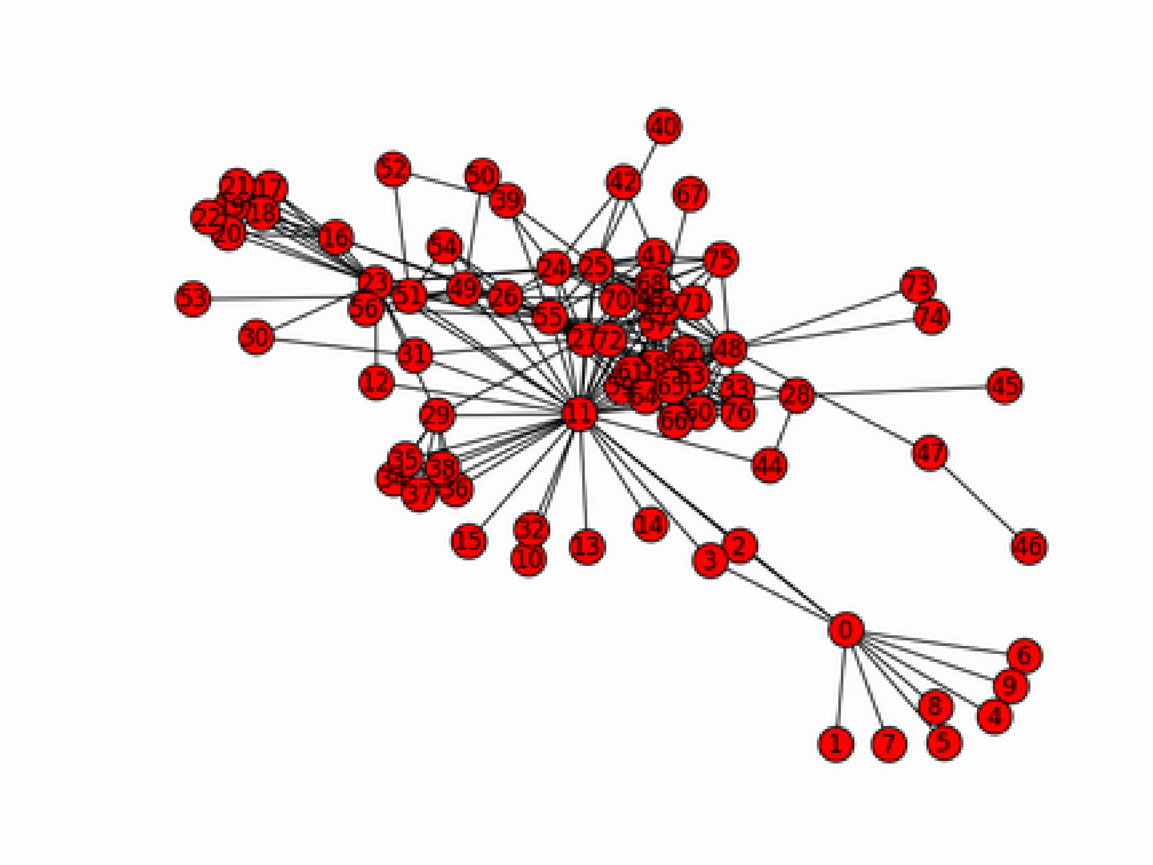

Fig 1. The Les Miserables Graph

[4] D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing*, Addison-Wesley, Reading, MA (1993)

First, I've just drawn the graph as it exists (Fig. 1), without the application of any algorithms. That said, it would be remiss of me not to mention that the very drawing of the graph is an algorithmic process rife with potential biases. How one draws a graph can implicitly force the viewer to think about it in a particular way, things in the center take on more importance for example, and a graph's layout is by no means automatically decided. Throughout this study I will be using a force-based algorithm (the Fruchterman-Reingold force-directed algorithm[5] specifically), that assigns forces to the edges as if they were springs, pushing them apart. The Les Mis graph has edge weights according to the number of coappearances of the two respective characters but these will not be used for the basic drawing process (unless mentioned explicitly). Each edge has equal weight.

So what can see from the graph as it currently stands? Looking at Fig. 1, without knowledge of the book or even the name of the person attached to each node, one node stands out. Node 11. From the graph one would suggest this is a crucial character in the drama and sure enough 11 is Jean Valjean, the closest thing to a main character Les Miserables has. The reason it pops out to such an extent is that at first appearance the graph has a hub and spoke type distribution, a good deal of the edges coming out from a central hub (node 11). So far then it would appear an graph approach is vindicated at least at an extremely basic level.

Centrality

Centrality comes in many forms but the essence of these algorithms is to try and find the relative "importance" of a node in a graph. What importance means is totally ambiguous and there are a whole host of different ways to define it. For all the following examples, the colour scheme will be the visual signifier of importance. The darker red, the more important (numerical data is also included in the attached documentation.

I'll examine four different centrality measures:

Degree Centrality, is defined as the number of edges incident upon a node (i.e., the number of edges that a node has to other nodes in the graph) otherwise known as the node's degree.

$$C_D(v) = \deg(v)$$

One interpretation of this is the likelihood of a node being used in a flow through a network. In the case of a directed network (where the edges are directional), there are two separate measures of degree centrality, indegree and outdegree. Indegree is a count of the number of links directed into the node and outdegree is the number of edges that the node directs to others. Our Les Mis graph is an undirected graph so the we are simply dealing with the number

[5] Graph Drawing by Force-directed Placement, SOFTWARE—PRACTICE AND EXPERIENCE, VOL. 21(1 1), 1129-1164 (NOVEMBER 1991), T. M. J. FRUCHTERMAN, E. M. REINGOLD

of edges to a node. From a narrative perspective having a high degree centrality means that many characters talk or meet with a given character. This would generally suggest that a character holds more importance, the more people they encounter the more significant they are to the story. That's the hypothesis at any rate!

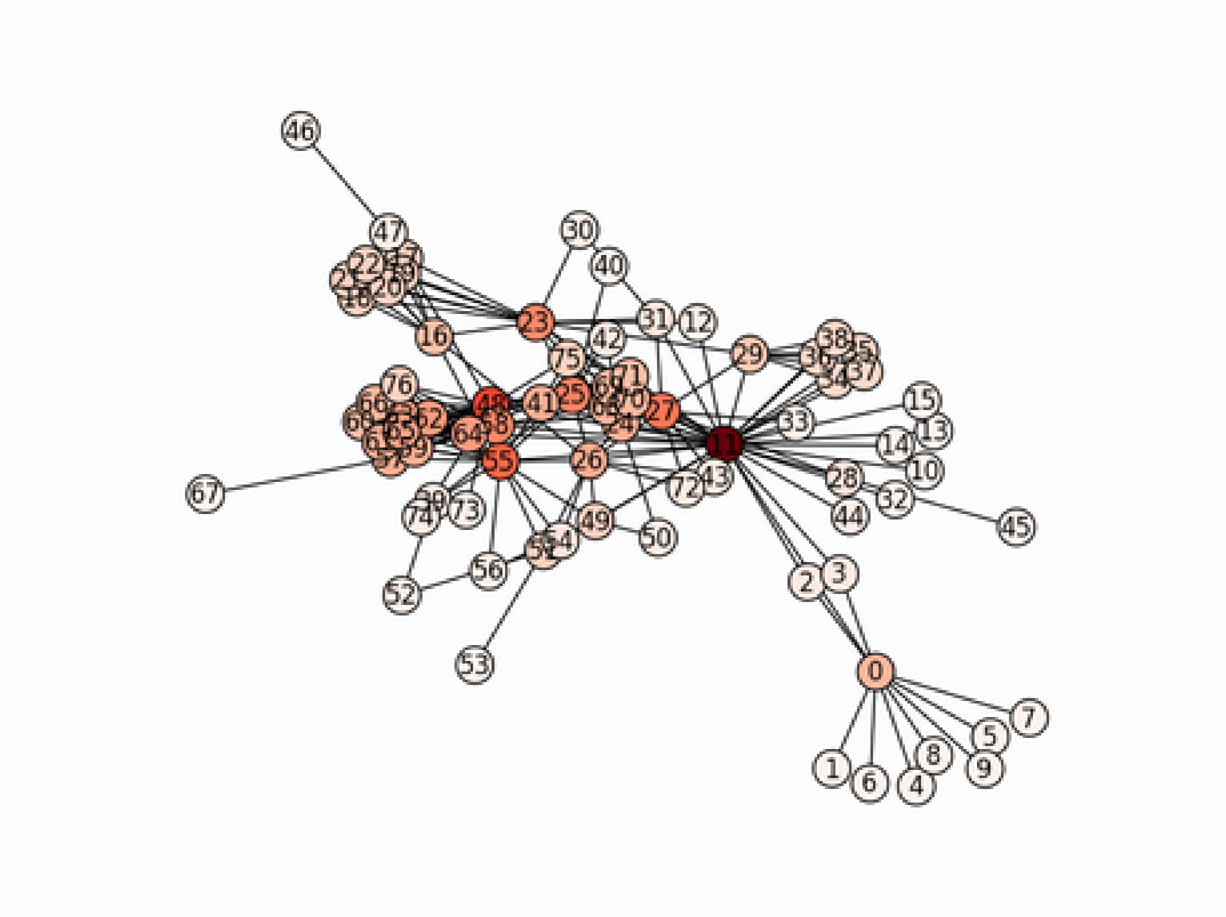

Fig 2. Degree Centrality of Les Mis Graph

So what do we learn from this type of centrality? Fig 2. reveals details that weren't apparent in the original graph. 11 still pops out so that's more evidence of the importance of that node but we also see 48 (Gavroche) emerging as well as 23 (Fantine), 25 (Thenardier), 27 (Javert) and 55 (Marius). 0 (Myriel) appears to be an outlier, being substantially detached from the rest of the graph.

The case of node 0 is an interesting one as in a larger graph, one that wasn't readily visualized, the fact that it clearly holds a more specialized position in the story might not be clear. Degree centrality gives us no information about relative position in the graph which could lead to inaccurate or ambiguous results.

Betweenness centrality defines the number of times a node acts as a bridge along the shortest path between two other nodes.

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Where $\sigma_{st}$is the total number of shortest paths between s and t and $\sigma_{st}(v)$is the number of these shortest paths that pass through v.

A high betweenness centrality therefore means that any information flow between two people would most efficiently go through this character. The argument from a narrative perspective for us is that high betweenness means that plot points, which must involve some degree of information transfer, are most likely to occur or involve characters who “connect” characters to each other.

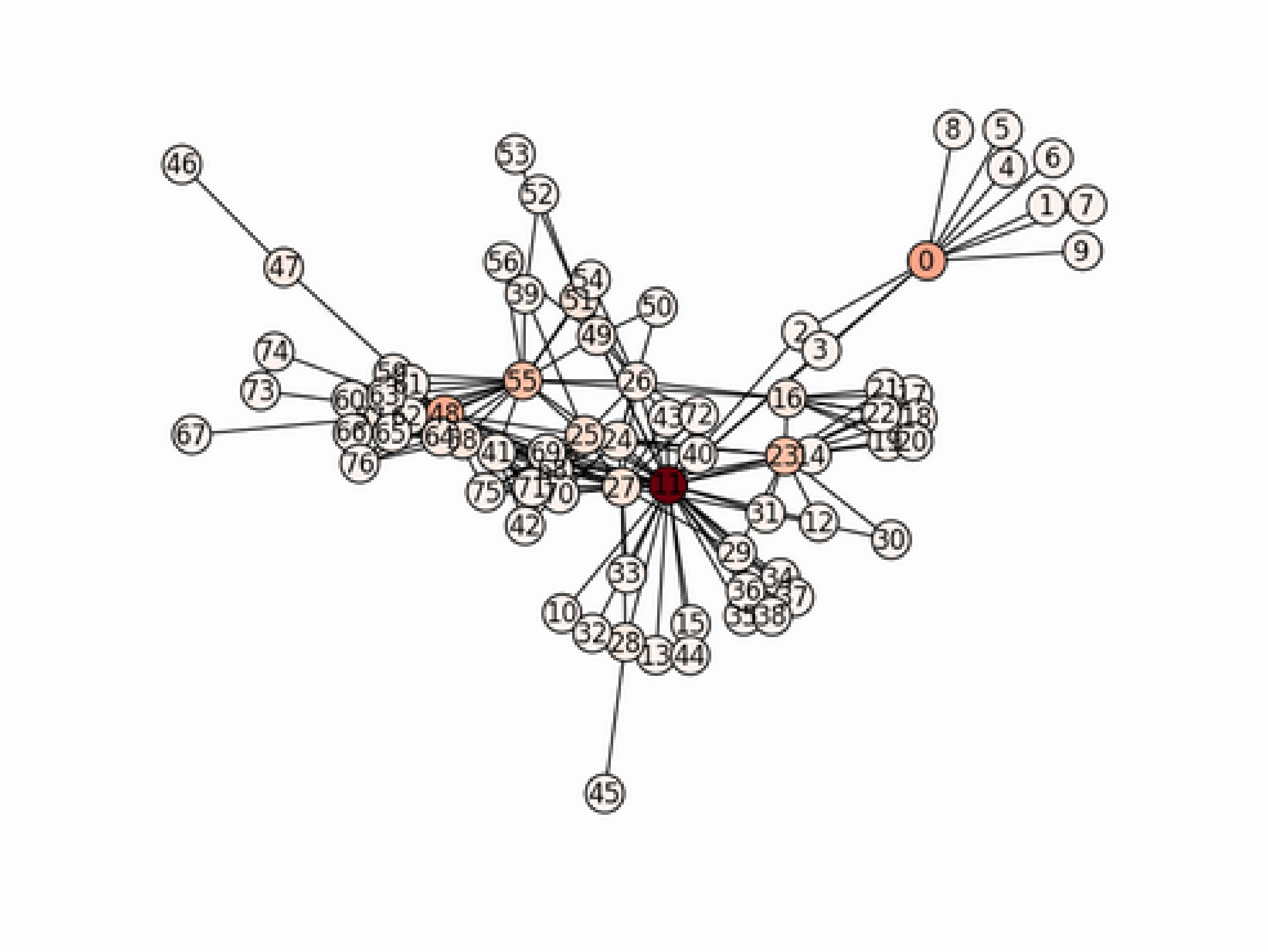

Fig 3. Betweenness Centrality Les Mis Graph

Familiar figures continue to jump out though Betweenness seems to have been more picky in which nodes are important. Unsurprisingly 11 does particularly well with this analysis, from our initial visual examination it was clear that 11 connects made of the separate parts of the graph.

Less obvious was how important 23 and 48 are relative to their neighbours whilst we can see again that 0 gets a high value here.

What we are beginning to see are a few patterns emerging. Betweenness is a very useful measure but, even more than degree, it gives potentially unfairly high scores to nodes that exists as the only route between one group of nodes and another. That could be significant, it might, for example represent a significant subplot, but in any application of this algorithm it is a potential problem. If we are not able to grasp that it represents the link from one subgraph to another, which isn't something shown in the undrawn data (looking at the pure values, 0 ranks second after 11), we might not be aware of its real significance.

Closeness Centrality measures the mean distance from a given node to every other node. This tends to mean central nodes (we'll look at another measure of "central" later) get higher scores as they are on average closer to every other node.

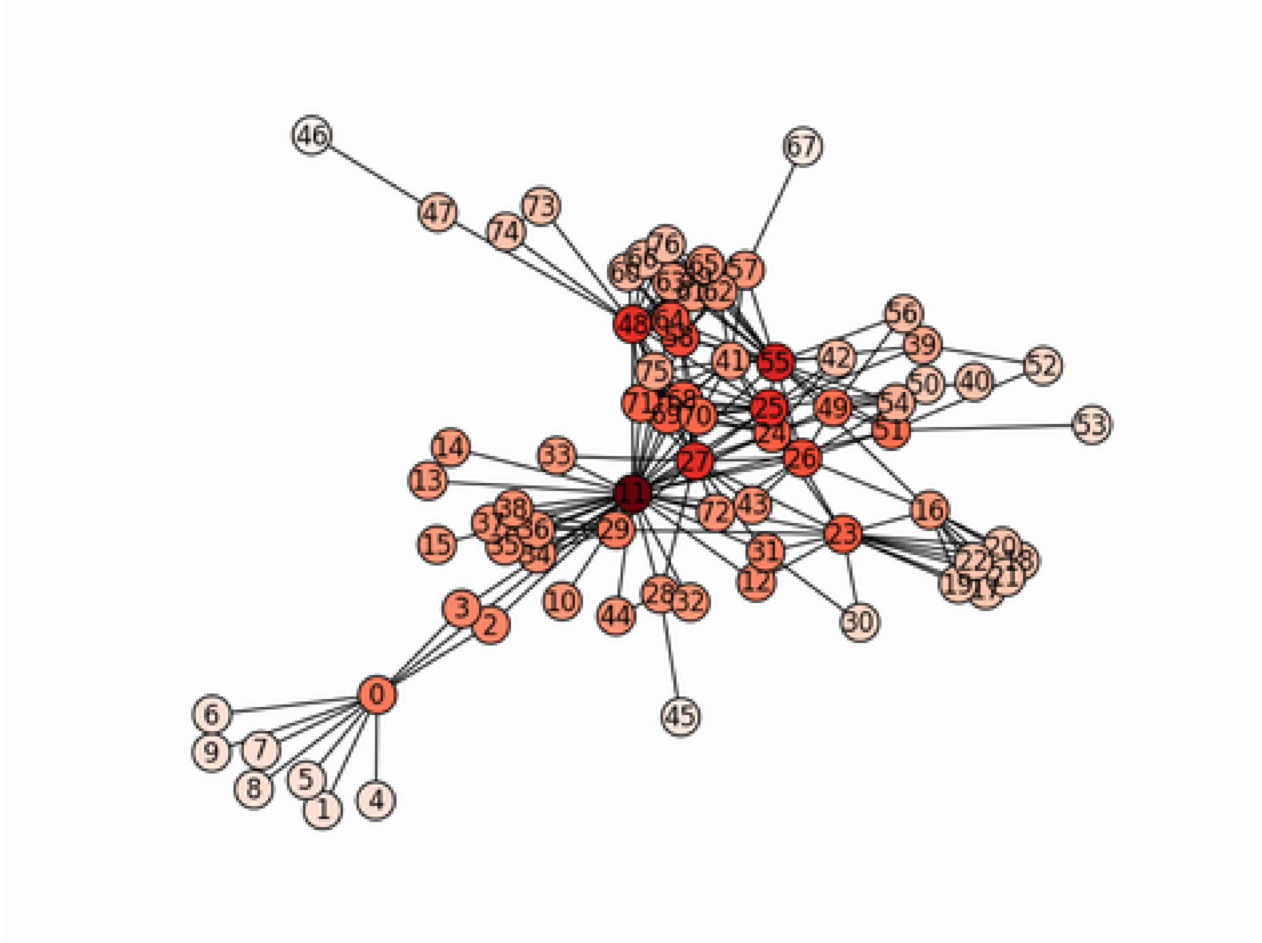

Fig 4. Closeness Centrality Les Mis Graph

This centrality places the nodes on more of a range than the other centralities, with less of the binary selection of some of the others. As a result it is harder to pick out a few individuals, but instead we get more of a spectrum of importance (Fig. 4). The strength of this measure appears to be how it confirms and negates the previous algorithms.

The top ten are the following:

```
11 (Valjean): 0.64406779661
55 (Marius): 0.531468531469
27 (Javert): 0.517006802721
25 (Thenardier): 0.517006802721
48 (Gavroche): 0.513513513514
58 (Enjolras): 0.481012658228
26 (Cosette): 0.477987421384
64 (Bossuet): 0.475
69 (Babet): 0.463414634146
68 (Gueulemer): 0.463414634146
```

A pretty solid list of the major characters (at least the first 7), though it’s intriguing that Babet Gueulemer get picked out so much more strongly here (2 of the 4 Patron Minette, the secondary villains for much of the book). Overall many of the same nodes still get the same importance but closeness picks out, at least to a certain extent, the lower importance of outliers like 0.

Eigenvector centrality is a measure of the influence of a node in a network. It assigns relative scores to all nodes in the network based on the concept that edges to high-scoring nodes contribute more to the score of the node in question than identical edges to low-scoring nodes.

The mathematics here is more complicated but in essence the algorithm runs iteratively from each node, working out the probability of hopping to another node from the current one at each stage. Given many iterations the values stabilize to what is mathematically known as the eigenvector for the matrix (the graph’s matrix representation) with the eigenvalues equal to the eigenvector centrality of each respective node.

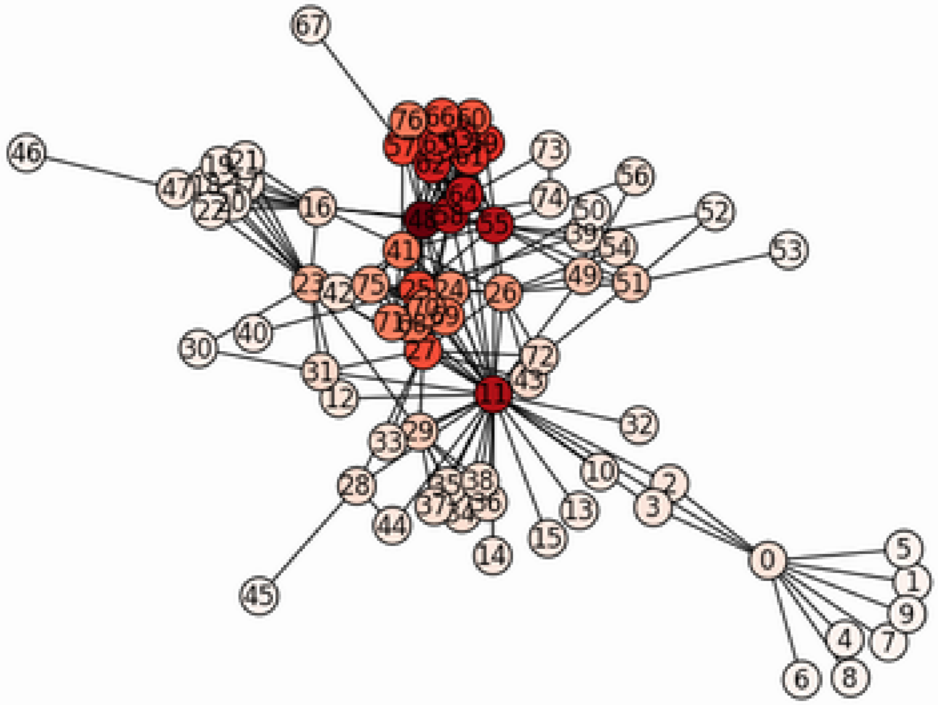

Fig 5. Eigenvalue Centrality for Les Mis Graph

We can seem some interesting aspects to this graph (Fig. 5), particularly in regards to how it differs from the other centrality measures. 11 still pops out (if it didn't I would be concerned) but 0 takes on far, far less importance. Particularly interesting is how the cluster numbered roughly in the 60s has been given such heavy weight. Strongly connected areas of the graph get the highest scores, as with many interlinking nodes the “hops” are likely to stay in the same area.

So what have we learned from all these centrality algorithms? It would seem from this case study that centrality is a solid measure for detecting important areas in the graph. It is also of note that different algorithms yield different results in terms of which nodes are considered important. One of the great strengths of this exercise however has been to show that in combination it is possible to verify the consistently important nodes. While many nodes have had more or less significance across the algorithms: 11 (Valjean), 55 (Marius) and 48 (Gavroche) have consistently ranked highly, which we can confirm are important characters in the narrative based on the actual text novel. Had we only taken one algorithm we might have been led to believe node 0 or the cluster around the 60s were significantly more important than they are, particularly if this had been a much larger data set so we could not readily visualize the connections.

Measuring the Graph

So there are many ways of analysing the importance of each node but we can also look at the graph as a whole. We can measure distances in a graph several ways but one of the key components of any analysis revolves around eccentricities. Eccentricity is defined as the maximum geodesic distance from any given node to any other node in the graph.

From these eccentricities we can find the diameter of the graph, which is the maximum eccentricity in the graph. A complete graph would have diameter one. A single string graph would have diameter equal to the number of nodes minus one. Our Les Mis graph has a diameter of 5, which means the longest chain in the graph contains 6 nodes (the graph as a whole contains 76). This reflects the relative sparsity of the Les Mis network with many nodes connected to only a few other nodes. The radius, which is defined in graph theory as the minimum eccentricity, is 3.

We can look for the nodes that have both these properties. Nodes with eccentricity equal to the radius are considered “centers” (A different concept from the centrality measures above). Perhaps surprisingly, our graph has a fair number of centers:

11, 25, 27, 48, 55, 58, 64, 68, 69, 71

A quick glance over this list reveals its resemblance to our Closeness centrality list. The only difference is 26 (Cosette) is removed and 71 (Montparnasse, another member of the Patron-Minette) added. This similarity suggests this isn’t a bad metric for finding important nodes but it falls down a touch because its resolution is low. 26 (Cosette) scores 4 and in the process joins the majority of the rest of the nodes in the graph. With a range of only 3 - 5 and integer values, for a graph of this relative sparsity, this form of measurement can only have limited benefits.

Looking at the other side of the coin, the periphery, nodes which have eccentricity equal to the diameter, we get:

1, 4, 5, 6, 7, 8, 9, 17, 18, 19, 20, 21, 22, 30, 45, 46, 50, 52, 53, 67

This is a greater number than is useful but it reflects something of the narrative structure of the book, which is to say, a vast number of characters are introduced, many of whom interact with only a few of the book’s more central protagonists (A narrative like 12 Angry Men[6] would have diameter 1).

Measuring the graph is unlikely to reveal any great narrative secrets then but in the case where the graph is huge it might at least give some reflections of how sparse the graph is. A graph with a high number of edges relative to nodes is liable to have a much lower diameter than the

[6] http://en.wikipedia.org/wiki/Twelve_Angry_Men

reverse (though not necessarily, and if we know the number of edges and nodes to compare this could be a point of interest). Such a dense graph would suggest the story spans many people in parallel rather than the more sectioned graph that we are dealing with here.

Spanning Trees

Another interesting group of algorithms concerns spanning trees. These are the minimal set of edges that completely connect a graph. These can be derived many ways, which edges are selected depends on the exact algorithm selected. Care must be taken as even the simplest variations such as between Prim (which will find a minimum spanning tree) and Dijkstra (which will find a tree of shortest paths from a source) can result in dramatically different graphs.

I'm going to consider Kruskal, a variation Min Span Tree algorithm, and for this we will need to consider edge weights. Our current graph edges assigned a higher weight for more frequent contact, i.e. more coappearances. A minimum spanning tree keeps the lowest weight edges so is not quite what we want. For this algorithm I am inverting all the values so that the largest become the smallest. We therefore keep the most important edges (here "important" meaning characters that most often come into contact) and lose the less important ones.

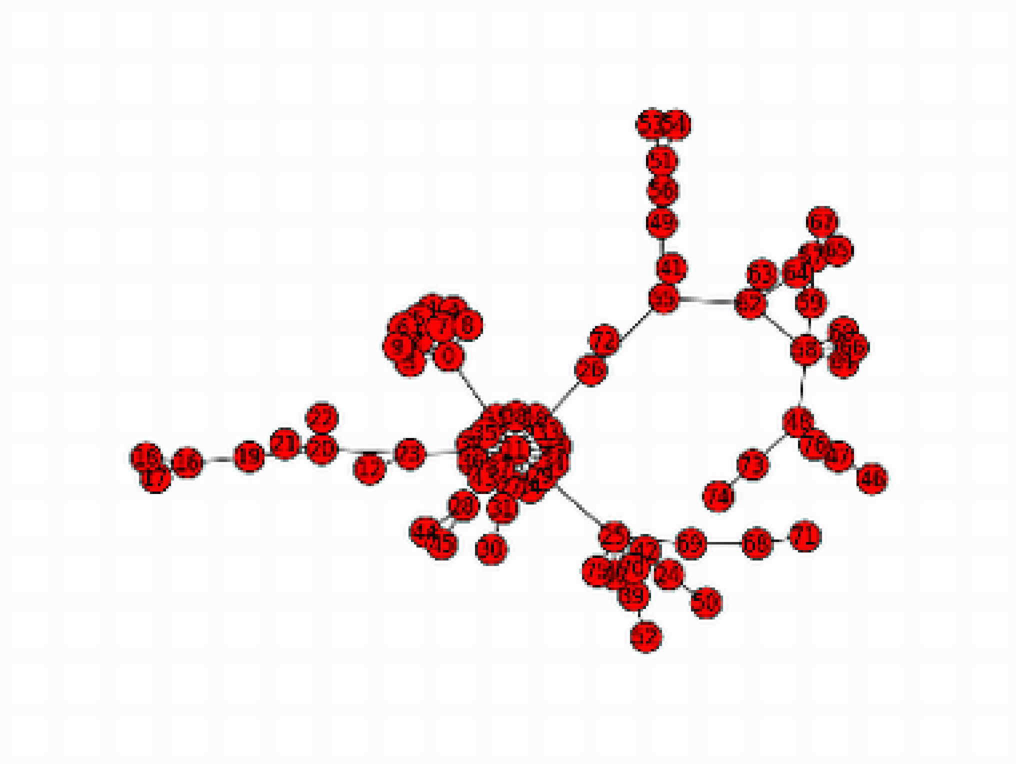

Fig 6. Minimum Spanning Tree for Les Mis Graph

Unlike the centrality measures we can only really derive stories from this graph by plotting it (Fig. 6), we aren't calculating flows so it is only when presented visually that we can work out the significance of the remaining edges. 11 yet again forms a tight, core hub but how the tree expands outwards is of note. A hypothesis that one might draw from this graph is that the sub trees such as the ones spanning from 23, 25 and 55 represent something like sub-narratives or characters who represent discrete of the overall tale. We have already learned from the earlier centrality analyses that few of the nodes further down the tree have overall significance but here we might be led to explore the 16 through 22 nodes impact on 23 or the nodes travelling up from 55.

In the case of the nodes moving left out of 23 (Fantine), looking up the names it becomes readily apparent that this is indeed a significant subplot in the narrative. These are all characters involved in Volume I, 16 (Tholomyes) 17 (Listolier) and 18 (Fameuil), the wealthy students who mistreat 23 (Fantine) and her friends 20 (Favourite), 21 (Dahlia) and 22 (Zephine). An impressive detection by the algorithm.

The edges moving up from 55 (Marius) prove less fruitful, perhaps on account of the more sprawling nature of the plot by this point. 41 (Eponine) fits nicely so close to Marius and the rest of the route covers a variety of quite closely connected sub-characters but it fails on some level because this graph was given an impossible task. Many of the characters who “should” have fallen in this subplot have fallen in others because of the immense amount of overlap between narratives.

The minimum spanning tree inevitably has the weakness that it only keeps the most important edges and therefore potentially loses information along the way. Each edge in the graph, even the weakest could be important from a narrative perspective and this is element is lost.

That said, the edges in this tree are importants one and there are no edges I could describe as wrong but by the nature of the algorithm we’ve lost the big picture. Unless it’s a situation like that of 23 (Fantine) or 0 (Myriel) where the narrative is almost completely disconnected from the overall narrative, this graph will only ever give us a partial picture.

Conclusion

Algorithms are no magic bullet, but they’re worth having in one’s armory. A journalist looking for a story must always make judgement calls and these algorithms do not circumvent that in any way, shape or form. Coming at this data set blind, these algorithms have proved useful and accurate, we know far more about the interrelationships of the characters and their relative importance now than before applying the algorithms. Centrality measures pointed towards the key players while the minimum spanning tree isolated several subplots but all these things are only pointers towards the truth, we needed research to back it up.

It’s also worth noting that quite apart from the difficulties encountered in the application of these algorithms, we are also dealing with only a very small part of the knowledge acquisition process. We’ve examined a graph with clearly defined nodes and edges but that is not always the case, rarely in fact, and even here we could have changed the definition of an edge (say to only characters who directly speak to another) and gotten a completely different graph. Between data acquisition, selection, cleaning and presenting[7] we've really only covered a single step and by no means the most difficult step.

Network algorithms haven't told us who lives, dies, falls in love or lives happily ever after (almost no one in Les Mis). They probably never will. But they can given us some key directions as to where to look to find these answers.

[7] Usama Fayyad et al.: *Knowledge Discovery and Data Mining: Towards a Unifying Framework*, KDD 1996.