Assignment #2 NodeXL

CMSC 734 – Information Visualization

Prof. Ben Shneiderman

Hua He

University of Maryland, College Park
Introduction

Classical novels are considered able to create accurate representations of real world, because typically novelists write novels based on real life experiences of their own. This idea is confirmed by many theorists and literature experts, who believe that certain aspects of novels, for examples, the plots, the dialogues or the relationships between characters, should generally follow real life patterns.

In this report, I would like to provide several insights on the patterns of characters' interactions and relationships within novels, based on social networks extracted from two 19th century novels.

Data

Two novels are prepared. The first novel is Les Misérables by Victor Hugo, I obtained its social networks in UCI social networks data base, this data set is originally from the Standford GraphBase by Donald Knuth[2]. The second novel is Pride and Prejudice (P&P) by Jane Austen, however, I cannot find other social network data sets online for other novels, so I have to extracted the social network of novel P&P by using my own program with the help of machine learning packages. The details of the automatic extraction method are described below briefly.

Automatic Social Network Extraction from Novels

To automatically extract social networks from novels is generally challenging, but if we can utilize results of speaker identification then this extraction process can be easier. Given an utterance (which is usually a quotes representing a conversation), what speaker identification can do is to determine who is the speaker of this utterance (select from all characters in this novel). My speaker identification program can directly utilize the surrounding linguistics features, and then uses a statistical SVM model to predict the most correct speaker for that utterance with high accuracy.

Once we can get results from the speaker identification program, we know how conversations in novels are formed, thus we can simply extract networks based on the conversations, for example, if A is talking to B, then A and B should be connected. The novel Les Misérables's social network extraction (not done by me) is different from the above speaker identification method; it is based on the number of such co-appearances within a chapter for any pairs of characters.

Both extracted social networks are good and precise enough to be studied and can reveal insights of characters' interactions within novels.

Novel Differences

For this report, I purposely choose these two novels to be in different categories so that the insights obtained from both novels are representative. Firstly, novel Les Misérables has more than 1000 pages while novel P&P is a much thinner book. Second, the stories in novel Les Misérables are more 'urban' (stories mostly in the city of Paris), while novel P&P is more 'rural' (farms and gardens). Third, they both have very different topics, for example, novel Pride and Prejudice is about love and marriage among several families (families seem to be units for P&P novels, Darcy family, Bennet family, Bingly family and others.)
Insights

'Connector' Characters Could Be Highly Important for Novel Storyline

Let's start with the thinner novel Pride and Prejudice first, its social networks are shown below,

Notice that each node represents a character, and links indicates the relationships between characters. Different colors are automatically determined by NodeXL graph clustering algorithm. This clustering algorithm can extract the closely connected components our of the whole networks and therefore is amazingly revealing. The colors can clearly distinguish different groups (families) of characters in this P&P novel, for example, the Darcy family is in dark blue, the Bennet family is in light blue, Gardiner family is usually with Bennet family and is in red, Bingley family is located between Bennet family and Darcy family and is in green.

Even without reading the whole P&P novel, we can easily tell that characters that are connecting two or more family blocks are very important, for example, the major character Mr. Darcy, the major character Jane Bennet, and the minor character Mr. Wickham. Among those connector characters,
Mr. Wickham is interesting. He is a seemingly good gentleman at first, but at the end Wickham turns out to be a Jane Bennet's nightmare. Over the whole P&P novel, the Darcy versus Wickham things are one of the major storyline, and Jane Bennet's attitude changes towards Mr. Wickham is also crucial. Interestingly, the network above can also shows that 'dangerous' triangle among Darcy, Jane and Wickham. Another example is Mr. Collins, who is again not a major character but is also a good player who can lead the storyline of the book to certain extent. Those minor characters that are purposely selected by the author to be the connector between the tight family blocks are important for the completely novel storyline.

**Having More Characters Does Not Lead to Less Conversational Patterns in Novels**

There is a literature theory on novels that was proposed by Fransco Moretti[3], that if a novel has more characters, then this novel's conversation part will tend to be smaller, and Fransco argues that social interactions of different kinds within novels will multiply, so that novel authors have to leave much less room for traditional conversational interactions among characters. This theory is debatable through years, and following work done [4], I would like to add more new evidences to argue against Fransco's idea on novels.

Now let's take a look at a more complicated and much bigger social network which is extracted from novel Les Misérables by Victor Hugo, with colors automatically generated by NodeXL's graph clustering algorithm:
If we follow Fransco Moretti’s theory, because novel Les Misérables has more characters, its conversational part should be smaller and therefore less connected. However, this is not the case for this novel. If we do a comparison between the above network from Les Misérables and P&P’s network, although Les Misérables has more characters (around 4X more in the network compared to P&P’s network), we can still clearly distinguish those closely connected ‘blocks’, for example, the Fantine block (in red) and the Bishop Myriel block (in oranges) are still closely connected. And the major character Jean Valjean (in the middle in dark blue) almost connects to all of its surrounding blocks.

The Conversational Patterns on Urban vs Rural Novels

Another interesting comparison in the literature theory is the urban vs rural setting. It is believed that novels with urban settings tend to have less conversational pattern, because characters are living in the city, which is complicated and could have many more interesting things for the characters to do than merely talking. However, this is still not the case given our examples. As shown below, apparently, novel P&P (on the left) is rural but has fewer connections compared to the urban novel Les Misérables (on the right). If we just count the connectness of both figures, both networks are closely connected, therefore the theory might not be correct.
NodeXL Critique

Strengths
1. NodeXL is useful to visualize various networks with little effort. It is easy to use, and all instructions on NodeXL are intuitive.

2. I like NodeXL also because its warning/error messages are helpful. Unlike many other applications, those messages from NodeXL can also provide solutions to solve users' problems.

3. The data input and export functionality is handy and productive.

4. The NodeXL program is doing a good job at providing feedback, this is following one of Eight Golden Rules [1]: Informative Feedback.

5. The consistency part of NodeXL is also good, seamless integrated with MS EXCEL, this is from Eight Golden Rules No.1: Consistency [1].

6. I love the 'groups' (e.g. graph clustering) functionality. Easy to use and display.

Suggestions
1. I guess others could mention this as well, the UNDO/REDO function seems to be not available.

2. I do hope someday there can be an animation feature added, that can 'display' the node linking process, for example, it can show users the process linking node 1 to 2, then linking node 3 to 1, etc.. And output the whole animation as an video.

3. Although NodeXL works only on Windows, I still prefer to have more short cuts

4. Seems to me NodeXL cannot do the full-screen display, which could be very useful for single screen users.

References

[1] Eight Golden Rules,  
http://faculty.washington.edu/jtenenbg/courses/360/f04/sessions/schneidermanGoldenRules.html  


[4] Extracting Social Networks from Literary Fiction, David K. Elson, Nicholas Dames, Kathleen R. McKeown