

# Lab Journal

Ross T. Harrison

Spring 2014

### **Abstract**

This is a journal of notes and experiments conducted as part of the requirements for ENGLISH 4/898 - SEC 002 *Characterization in Literature: a Macroanalysis*

## 0.1 Structural Analysis on Social Networks Constructed In Literature Texts(Park)

January 18, 2014

### 0.1.1 Goals

Use techniques applied to economics, sociology, and other disciplines on literature. The purpose is to measure connectivity between characters.

Previous work has investigated social structure in novels as a static whole (Elson). Their eventual goal is to investigate the change in social structure over time. The goal of this paper is to demonstrate the effectiveness of their co-referencing technique

### 0.1.2 Technique

Their technique is similar to previous work. Using the textual distance between mentions of a character to calculate the weight of their connection. A key difference with their algorithm is they use distance in statements instead of words. That is if two characters are in the same sentence then their distance is 0. If in adjacent sentences their distance is 1.

The weight is then calculated from the distance using a power function. This is nice because it has a maximum value (1) and rapidly (depending on variable) approaches 0, but never reaches. The relationships then have a maximum weight, but are never "meaningless". For computational purposes however, the group decided to enforce an influence region restriction, or how many statements ahead will the algorithm look. They used 10 as their influence region.

### 0.1.3 Conclusions

Using their work as a guide could be useful, but the paper is too poorly written, ill defined, and data delivered in to low a resolution, to investigate their exact results.

However, their idea of investigating changes in networks over time is interesting. However, if I were to pursue it, I would like to create a "real time" viewer that would allow the user to scrub the piece, and see the relationships change fluidly.

## 0.2 Character Social Networks

January 26, 2014

A recurring idea is that character archetype can at least in part be determined by relations with other characters. This problem can be broken into several subproblems, first would be determining character existence, second the

existence of a relationship between those characters, and thirdly the nature of that relationship.

The first subproblem is the most quantifiable. Since for many novels a complete, or semi-complete list of characters is available. Success metrics are therefore readily available. The second and third seem more difficult. Defining a correct answer for character relationship existence, much less quality, is much more difficult, if not intractable.

Relationship existence at least seems to have a possible solution, comparing with human created character networks. This has the obvious drawback of being extremely labor intensive in constructing a test set. Therefore, unless such a corpus is readily available, a prudent course of action might be to investigate network creation in a more qualitative manner.

## 0.3 Character Network Construction

: (January 26, 2014)

Character networks can be constructed using dialogue (Elson, Dames McKewon) Or by simple coocurrence in text (Park, or LOTR project)

Like Park et-al, I'm interested in investigating social changes over time. However, while they intend to divide the text into chapters, I believe that a more interesting view would be to construct a social graph for any given point in the text.

My idea is create a web interface that allows the user to upload text and a character list. Then produces a network similar to the one seen on the LOTR project. However, this graph will allow the user to adjust constants.

The user will also be able to "scrub through the text" viewing the changes in real time. Using statements as the atomic units, the connections made would be further weighted by distance from current position using some kernel. A variety of options could be available: \* A power function \* a normal function

The idea then is to see in "real time" the changes in importance and strength of character relationships.

Exposing the constants and filters to the end users

## 0.4 Elson

January 27, 2014

This article was referenced by park. It is looking at two claims made about 19th century british novels:

1. there is an inverse correlation between coversation and # of chars
2. There are more interactions occuring in rural, rather than urban settings

They intend ot construct networks based on dialogue. First finding quoted speech and attributing to a set of characters. Edge weights are determined by the ammount of conversation going on between the characters.

60 novels, 31 authors

They claim that this study contradicts widely held notions

#### 0.4.1 Related Works

Burrows 2004, word usage patterns. Mostellar and Wallace. “outing” authors  
Lee, 2007. Validating lineage of ancient texts

Semantic analysis has been more rare. Exceptions: Chambers and Jurafsky  
2008

Moretti (2005) graphically mapped out text according to: \* geography, \*  
social connection \* etc

#### 0.4.2 Hypotheses

Statistical methods are essential for testing validity of core theories.

Bakhtin 1981, 84. Different spatial settings have different potentialities and govern social interactions in a way that should turn up in analysis. -> Rural communities are smaller tighter knit groups than urban. Moretti (1999) argues, horizontal connection are more important in city than transgeneration. As the number of characters increases the social connection become more complex and the whole system becomes unstable, and blur lines between character roles.

Novels are divided into urban, rural, and mixed groups. Presumably manually. They were selected as representative based on canonical authors, decade, setting, and sub genre.

- Urban - set in metro, multiple labor forms.
- Rural - set in village, agriculture is primary, landowning gentry are prominent.

Pulled from Guttenburg

#### Hypotheses stated

1. Inverse Correlation Between # of characters and the amount of dialogue
2. 19th cent brit lit. depicts urban groups and large and loosely related. rural as tightly bound smaller.

These two hypotheses are potentially related, as there are generally more people in cities.

#### 0.4.3 Extracting Conversation Networks

Conversation:

1. continuous span of narrative
2. characters in same place at same time

3. characters take turns speaking
4. characters are mutually
5. aware of each other's speech, and are intended to hear.

### **Character Identification**

First each novel was processed with Stanfor Named Entity Recognition tagger (Finkle 2005). Noun phrases extracted that were categorized as person or organization. Nouns clustered into entities.

Clustering: 1. For each Named entity. 1. Generate variants \* Mr. Sherlock Holmes, Mr. Holmes, Sherlock Holmes, etc 2. Compile list of coreferent names, variations etc 2. Aggregate mentions

Text was also preprocess to normalize formatting and detect sections, especially looking for quoted speech.

### **Quoted Speech Attribution**

Approach Described in previous workd (Elson & McKeown, 2010). Training set of Brit, American and Russian Texts. Amazon Mechanical Turk provided human text cases. Accuracy was like 83%, but because conversation is what is desired, not quotes, success is likely higher. Precision emphasized over recall. That is the emphasis is on the percentage of valid answers not on retrieving all valid answers.

## **0.4.4 Evaluation**

Four novels were held out of the training set. And only used to evaluate. For each novel random chapters were selected, manually processed. Multiple conversations were counted as individual components.

Speech Adjacency has high precision (.95) recall 0.51 (almost half the results were missing). correlation and spoken mention methods (alternative edge creation methods) had much lower precisions (0.21, 0.45), but similar recall (0.21, 0.45).

## **0.4.5 Data Analysis**

### **Feature Extraction**

1. Number of characters and Number of speaking characters (separate features?)
2. The variance of the distribution of quoted text among n most frequent features for n between 1 and 5
3. Number of quotes and percentage of novel as quoted speech
4. Number of 3-cliques and 4-cliques in social network

5. Average Degree of the graph, that is the average number of edges per node
6. Graph density, normalizing the average degree by number of characters:
  2. The idea is to capture what percent of the entire network does each character average.

Weight of edge over 0 doesn't affect feature 5 or 5.

## Results

Pearson's product-moment correlation.

Hypothesis 1 - They found a weak positive correlation. Stronger correlation between number of unique speakers and number of quotes. Also, the connectedness (average degree) of the graph had a positive correlation with # of chars. Analysis suggests the opposite is true, small communities tend to be more disconnected.

Hypothesis #2 - No significant difference in the size of the graphs. Simply "not confirmed"

However, third person narrative, does have more frequent connectivity. This makes sense since monologue isn't really a huge deal.

Aside, it is interesting to think about ASOIF in this case, since it seems to mix first and third person. The character perspectives shift, and language is third person, the reader is bound by the characters perspective and thoughts.

### 0.4.6 Literary Conclusion

Narrative Voice trumps setting.

### 0.4.7 conclusion

High precision for detecting face to face communication between two named characters. Narrative perspective is a much stronger predictor of graph characteristics

## 0.5 Stanford Papers (pulled from their NLP site)

### 0.5.1 Baselines and Bigrams: Simple, Good Sentiment and Topic Classification

January 30, 2014

Investigating performance and efficiency of Naive Bayes (NB), SVM, and Hybrid solutions on Sentiment tasks. NB actually does better. Hybrid can provide new state of the art performance level.

Data set includes customer reviews, movie reviews, IMDB, newsgroups.

Tokenization was used when available. If not unigrams were created. Non Alphabetical were filtered out. Cross validations splites were observed in several cases. Delta values for  $p \leq 0.5$  were displayed.

MNB better at snippets. Statistical methods miss some cases in smaller sets. “Not An Inhumane Monster”, “Killing Cancer”.

SVM is better at full-length reviews. When the target of sentiment in a larger body of text then, SVM performs better.

NVSVM especially bigram features performed well across the board, large and short sections. “Often gives better results than previously published” (trigrams hurts slightly).

Utility? See what features are being used for sentiment analysis when it comes to it. Need to learn more about what features are in the SVMs in general.