The Sense and Sensibility of Different Sliding Windows in Constructing Co-occurrence Networks from Literature

Siobhán Grayson[1], Karen Wade[2], Gerardine Meaney[2], and Derek Greene[1]

[1] School of Computer Science, University College Dublin, Ireland
{siobhan.grayson, derek.greene}@insight-centre.org
[2] Humanities Institute, University College Dublin, Ireland
{karen.wade, gerardine.meaney}@ucd.ie

Abstract. In this paper, we explore the design and effects of applying different sliding window methodologies to capture character co-occurrences within literature in order to build social networks. In particular, we focus our analysis on several works of 19th century fiction by Jane Austen and Charles Dickens. We define three different sliding window techniques that can be applied: collinear, coplanar, and combination. Through simple statistical analysis of each novel's underlying textual properties we derive tailored window sizes for each case. We find that the selection of such parameters can significantly affect the underlying structure of the resulting networks, demonstrated through the application of different social network metrics on each of our novels.

1 Introduction

Computational approaches are being increasingly adopted by humanities scholars to explore questions in the field of literature from new perspectives [6]. In particular, social network analysis (SNA) provides researchers with an array of existing analysis techniques, together with a unique level of abstraction (*i.e.* a network of nodes and edges), whilst still maintaining the social structure of novels and the societies they depict. The application of SNA in a literary context often involves the construction of *character networks* from a digital text, where each node in the network represents a character and each edge indicates some kind of relation between characters. Using these networks, methods from SNA potentially allow humanities scholars to test existing or new literary hypotheses from a quantitative perspective, in conjunction with existing close reading strategies. Unlike when dealing with modern texts, for 19th century fiction the extraction of networks is non-trivial, where text formatting may be inconsistent or lost and many characters are referred to by the same first name or surname. To date, most literary social networks have been extracted automatically, with authors making allowances for their incompleteness or inconsistencies [2–4].

In this paper, we describe three different *character network* construction strategies to detect co-occurrences, based on sliding a window over the text of

each chapter in a novel. Co-occurrences are then used to construct a weighted undirected social network for the novel. We demonstrate the impact of the choice of method and associated window size parameter using nine popular 19th century novels written by the British authors Jane Austen and Charles Dickens, available from Project Gutenberg. These texts have been manually annotated in order to include as many character entities as possible, including minor and collective-presenting characters. We illustrate how altering the network construction method affects the structure and density of the resulting character networks.

2 Related Work

A range of different approaches have been considered to identify meaningful interactions between characters in fictional texts. Moretti [6] analysed the works of Shakespeare by constructing networks defined on the basis of dialogue alone. However, when dealing with prose, limiting interactions to quoted speech will exclude large amounts of non-quoted dialogue, observations, and thoughts [1]. Elson *et al.* [3] constructs social networks from 19th century literature by detecting conversations from sets of dialogue acts, which involves character name clustering followed by automated speech attribution. While this approach achieves a high level of precision (96%), the level of recall for conversational interactions is low (57%), even before other types of character interactions are considered. In an attempt to overcome the limitations of using dialogue alone, Agarwal *et al.* [1] examine two distinct types of social events involving characters in Lewis Carroll's *Alice in Wonderland* (1865): interactions and observations. The authors construct a weighted undirected social network from instances of the former, and a weighted directed network from instances of the latter where edge direction is based on who is observing whom. More recently, Jayannavar *et al.* [4] also apply an extraction technique which goes beyond dialogue, looking at the network of general character interactions, as well as considering specific cases of conversational interactions and observations.

Beyond the study of literature, co-occurrence analysis has often been used to identify the linkages between words in unstructured texts. For instance, the relationship between pairs of terms occurring within a constant-sized context window is a key component of popular word embedding methods such as *word2vec* [5]. In topic modeling, the frequent co-occurrence of a pair of terms within a sliding window of fixed size moving over a corpus is used to measure topic coherence [7]. In both applications, the choice of context window size is often not considered in detail. However, Zadeh and Handschuh [8] demonstrated the importance of context window sizes when identifying co-occurring terms for the purpose of classification.

3 Methods

Data Preparation. We consider a collection of nine novels from two 19th century British novelists - six by Jane Austen and three by Charles Dickens -

sourced from Project Gutenberg. Initial data preparation involves the manual annotation of the novels, where literary scholars identify all character references in the text of each novel. The annotation process itself consists of a number of steps. Firstly, a *character dictionary* is constructed, which includes a single entry for each unique character in the novel (identified by their *definitive name*) and the corresponding *aliases* for that character which appear in that novel. Once the dictionary has been compiled, all instances of a character's aliases in the novel text are replaced with their definitive name. To construct a network, a node is created for each character in the novel's character dictionary. Each chapter of the annotated text is then tokenised and an appropriate strategy is applied to identify and count all co-occurrences of character mentions. We then create a weighted character network for the chapter, where edges are weighted to reflect multiple co-occurrences. Finally, we construct an overall network for the novel by aggregating the individual networks from all chapters.

Collinear Co-occurrence Window Strategy. In this strategy, a sliding window of size w_l tokens moves over the text of each chapter. A co-occurrence between characters X and Y is identified when Y appears after X within this window. The strategy is collinear in that only consecutive pairs of characters are counted, and it is conservative in the sense that a co-occurrence between Y and another character appearing prior to X is not counted. This can be viewed as a variant of the left-hand context window approach described for term co-occurrence in [8]. The size of the sliding window w_l is identified independently for each novel. Firstly, we construct an overall character network for each window size $w_l \in [20, 300]$ words. We then calculate the weighted edge density as w_l increases and plot these values. Finally, we automatically identify the point at which this plot plateaus. This indicates that increasing the window size further will not capture any additional unique character interactions. Details of the resulting window sizes are provided in Table 1.

Coplanar Co-occurrence Window Strategy. Our second strategy is less conservative in that it aims to capture associations beyond pairs of consecutive mentions as illustrated in Fig. 1 (b). Due to the nature of coplanar connections, the method used to derive collinear window sizes is not applicable. This is because as the window size increases, rather than plateauing, the weighted edge density continues to increase until every character is connected to each other. Instead, the number of tokens between characters, referred to as "gaps", are analysed. The theory being that as the number of tokens increases between characters, the probability of an interaction decreases. Thus, treating

(a) Collinear Links **(b)** Coplanar Links

Fig. 1: Example of different window techniques demonstrated using an except of text from Chapter 2 of *Bleak House* by Charles Dickens.

Novel	N	$\#T$	$\#Chap$	w_l	w_{p1}	w_{p2}	w_{p3}
Northanger Abbey	94	75153	31	130	45	72	99
Pride and Prejudice	117	120262	61	90	34	54	74
Persuasion	136	81809	24	90	37	60	83
Sense and Sensibility	158	118149	50	70	36	57	78
Emma	193	156364	55	100	37	59	80
Mansfield Park	218	157800	48	90	39	62	85
Oliver Twist	286	153990	53	120	32	51	69
Great Expectations	288	177043	59	110	39	63	87
Bleak House	516	341441	67	100	36	58	79

Table 1: Summary of overall character network properties for the novels in our study (6 from Austen, 3 from Dickens) and selected window sizes. Here N is number of characters, $\#T$ is number of tokens (including character mentions), w_l is the collinear window size, w_{p1}, w_{p2}, w_3 are coplanar window sizes. All window sizes are in unit tokens.

gaps as the boundaries of character interaction events, window sizes are generated by exploring the most probable upper limits derived by applying simple, non-parametric statistical analysis on each text's gaps distribution (D_g). In particular, we take advantage of the interquartile range ($ICR = Q_3 - Q_1$) to define $inf(D_g) = Q_1 - 1.5 \times IQR$ and $sup(D_g) = Q_3 + 1.5 \times IQR$ where Q_1 is the first quartile, and Q_3 is the third quartile. Any elements which lie outside these limits are considered suspected outliers and are trimmed. Three window sizes are then considered: $w_{p1} = Q_3$, $w_{p2} = (sup(D_g) + Q_3)/2$, and $w_{p3} = sup(D_g)$.

Combined Sliding Window Strategy. As described above, the coplanar strategy captures associations beyond pairs of consecutive mentions, however, this is at the expense of seizing potential interactions which are further spaced out, and which would be naturally accommodated for by the larger window sizes enjoyed by collinear methods. Thus, the combined strategy consists of executing both the collinear and coplanar methods to identify character interaction pairs. The resulting co-occurrence pair sets are then merged, where pairs present in the collinear method, but not the coplanar, are added to the coplanar pair set.

4 Results

A summary of each novel's properties and the resulting window sizes is given in Table 1. Interestingly, Austen's *Northanger Abbey* has the largest collinear sliding window with $w_l = 130$ despite having the least amount of tokens $T = 57153$. It also has the highest coplanar window sizes indicating that a larger amount of text passes between character mentions within the plot. However, this correlation is not observed elsewhere, for instance, *Oliver Twist* has the second highest collinear window size ($w_l = 120$) but generated the lowest coplanar window sizes. To quantify the effect of each window strategy in terms of network topology, we have applied a number of common SNA metrics which we now discuss.

Network Analysis. As expected, each graph's weighted edge density, d_w, increase as we move from collinear, to coplanar, through to combination. Demonstrating how the collinear method is more robust against reaching the upper limits of graph density and how a high density is a natural consequence of the coplanar strategy. We also measured the average node disconnect within each graph and found it decreases from collinear to coplanar and combination. In Fig. 2, the overall network of *Oliver Twist* is visualised for each window strategy where the same group of four characters have been highlighted and focused

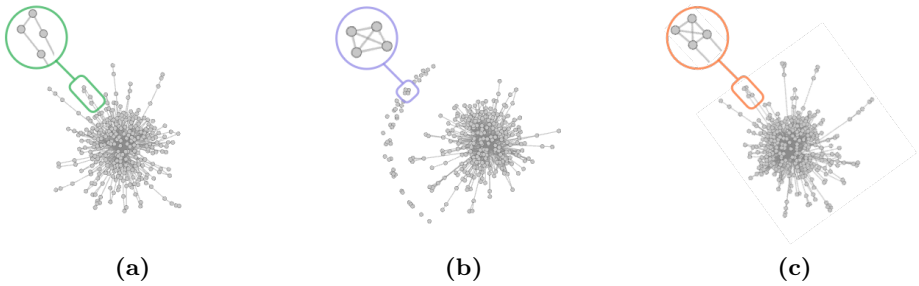

Fig. 2: Overall network of *Oliver Twist* with the same group of four characters highlighted and focused on in each case where (a) is collinear, (b) is coplanar $w = 32$, and (c) is combination $w_l = 120, w_p = 32$.

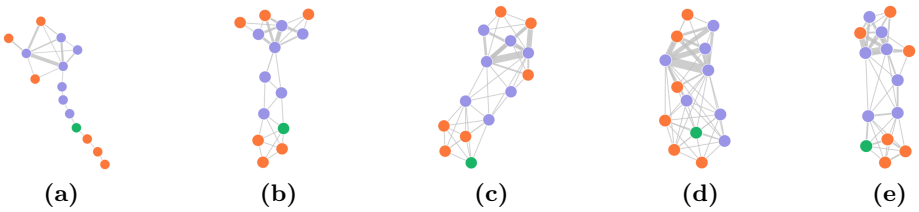

Fig. 3: Chapter 12 networks from Austen's *Pride and Prejudice* using four different sliding windows. (a) is collinear, (b,c,d) are coplanar, and (e) is combination with $w_p = 54$. Nodes coloured according to gender, purple is female, orange is male, and green is NA.

on in each case. Fig. 2 (a) represents the collinear network which on closer inspection shows the group of four character interactions occurring in a chain. Fig. 2 (b), depicts the coplanar-32 network. In stark contrast, there are a large number of disconnected characters and groups, including the previous group of four. However, it has established interactions between the members themselves. Finally, Fig. 2 (c) reconciles both approaches showing the resulting combination-32 network where not only are associations between characters within the group preserved, but they are also attached to the remainder of the network through links originally established by the collinear approach.

Another way of illustrating the effects of each window strategy is to compare the average clustering coefficient (C) and average betweenness (B) of all characters. We found both C and B decrease as we move from collinear to coplanar through to the combination. These results highlight how the collinear strategy primarily forms edges in a chaining succession (see Fig 3 (a)), causing the same characters within chapter 12 of *Pride and Prejudice* to be linked in such a manner as to have inflated clustering and betweenness values in comparison to coplanar (Fig 3 (b,c,d)) and combinational models (Fig 3 (d)) for the same text.

Discussion. To examine the effect of using different window strategies and sizes on the character rankings, we focus on *Oliver Twist* by Charles Dickens. The difference is quickly apparent when we consider only the top five characters ranked by degree. Strikingly, not one of the coplanar networks replicates the ordering of

the five characters. Highlighting the influence window size can have even on "major" nodes within a network. Interestingly, collinear ($w_l = 120$) and coplanar-32 are most comparable, despite their completely different methodologies and sizes. When extended to view the top ten degree characters of *Oliver Twist* not only do characters change ranking but there can be differences in the characters that appear too. For instance, The Artful Dodger replaces Mrs. Maylie within the top ten degree ranking for the collinear network, and supplants Nancy from the top ten degree ranking within the coplanar-32 network.

5 Conclusions and Future Work

In this paper, we have presented three different sliding window strategies that can be employed to capture character associations and generate character networks from potentially poorly-formatted literary texts. Our findings suggest that the choice of strategy is non-trivial, and can have a considerable impact on the resulting character networks. To ascertain the suitability of each, the analysis of the resulting networks should not be carried out in isolation, but performed in the context of the original texts themselves with the desired associations in mind. As a next step, we plan to compare the results of the different strategies against a gold standard of hand-annotated interactions that relate to our radically inclusive annotation methodology and the associated hypotheses relating to gender, genre and the nationality of the author.

Acknowledgments. This research was partly supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, in collaboration with the Nation, Genre and Gender project funded by the Irish Research Council.

References

1. A. Agarwal, A. Corvalan, J. Jensen, and O. Rambow. Social network analysis of Alice in Wonderland. In *Proc. Workshop on Comp. Linguistics for Literature*, pages 88–96, 2012.
2. A. Agarwal, O. Rambow, and R. J. Passonneau. Annotation scheme for social network extraction from text. In *Proc. 4th Linguistics Annotation Workshop*, pages 20–28, 2010.
3. D. K. Elson, N. Dames, and K. R. McKeown. Extracting social networks from literary fiction. In *Proc. 48th Meeting of Assoc. Comp. Ling.*, pages 138–147, 2010.
4. P. A. Jayannavar, A. Agarwal, M. Ju, and O. Rambo. Validating literary theories using automatic social network extraction. *Proc. 4th Workshop on Comp. Linguistics for Literature*, pages 32–41, 2015.
5. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.
6. F. Moretti. Network Theory, Plot Analysis. *New Left Review*, 68:80–102, 2011.
7. M. Röder, A. Both, and A. Hinneburg. Exploring the space of topic coherence measures. In *Proc. 8th Int. Conf. Web Search & Data Mining*, pages 399–408, 2015.
8. B. Q. Zadeh and S. Handschuh. Evaluation of technology term recognition with random indexing. In *Proc. 9th Int. Conf. on Language Resources and Evaluation*, pages 4027–4032, 2014.