Temporal Network Analysis as a Method to Explore World War I Diaries

Ashley Dennis-Henderson, Matthew Roughan, Jonathon Tuke

Abstract for GrapHNR 2023

Background

World War I was a significant event in Australian history which has been extensively researched. One aspect of this research is the analysis of war diaries. This analysis usually involves close reading of the text, and as such, the number of diaries that can be analysed in a single project is generally small.

In recent years, the State Library of New South Wales has digitised and transcribed a collection of approximately 550 Australian World War I diaries. This paper explores whether network analysis can be used to find diaries (or diary entries) within this collection which should be analysed using close reading.

More specifically, our research questions are:

1. Can network analysis be used to find days of interest which should be further explored by close reading?
2. Can network analysis be used to find diaries which are representative of this collection?

Methods

Dates are initially extracted from the diaries and cleaned using a combination of regular expressions and optimisation. These dates are then used to divide the diaries into dated entries. Using our dated entries, a temporal network is created, where nodes are individual diarists. On each day during the war, two nodes are connected by an edge if both diarists have entries for that day and those entries have high similarity. To determine similarity, the entries are converted to a vector representation using BERT [1], and the cosine similarity between these vectors is calculated [2]. Similarity scores range from -1 to 1 (-1 is complete opposite, 0 is no similarity, and 1 is complete similarity). Edges are only included in this network if their similarity is in the highest 10% of similarity scores. This network is then considered using two different methods.

Firstly, dates of interest are found by considering the dates in our temporal network with the highest number of edges. The static network for each date of interest is then considered, with the sentiment of each entry (calculated using VADER [3]) assigned as a node attribute. If this static network is large, communities will be detected to aid analysis.

Secondly, this temporal network is converted into a static network where each node is a diarist, and edges exist between two nodes if those diarists had similar entries on at least one day. The edge weight is the number of days in which those diarists had similar entries. The degree and weighted degree of each node is then considered. A high degree indicates a diarist whose entries are similar to some entries from many other diarists. A high weighted degree indicates a diarist whose entries are similar to many other entries. Therefore, a diarist with both high degree and high weighted degree should be representative of this collection. Preliminary analysis on a small subset of diaries (n = 20) has been performed to explore these methods.

Findings

For this subset, it was found that 25th April 1915 had the highest number of edges, followed by 20th May 1915 and 24th May 1915. Figure 1a shows the static network for 25th April, which has a primary group of five diarists, four with negative sentiments (1, 10, 15, 18), and one with positive sentiment (3). Close reading shows that all five diarists were writing about the Gallipoli landings. The diarists with negative sentiments all went ashore that day, were fired upon, and saw countless men wounded and killed. Whilst the diarist with positive sentiment was also involved in the landings, they did not go ashore, and talked about the good news they had heard regarding the progress of the battle. On 20th May and 24th May there is a primary group of 5 connected diarists discussing heavy fighting and a temporary armistice, respectively. This shows that this method has the potential to find interesting entries for close reading and should be investigated on the full set of diaries.

To test our second method, the static network of diarists was created (Figure 1b). In this network it was found that diarist 16 had the highest degree (18), followed by diarist 14 (17). Whilst diarist 14 had the highest weighted degree (204), followed by diarist 16 (136). This would suggest that these two diarists would provide a good overview of the diaries in this set. This is in line with our expectations as these diaries have the largest number of words and cover the entire period of the war. Furthermore, these diarists cover many aspects of Australia's involvement in the war (e.g. Papua New Guinea, Gallipoli, Western Front).

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, 2019, pp. 4171–4186.

[2] J. Han, M. Kamber, and J. Pei, "Getting to Know Your Data," in Data Mining, 3rd ed., J. Han, M. Kamber, and J. Pei, Eds. Boston: Morgan Kaufmann, 2012, pp. 39–82.

[3] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, no. 1, pp. 216–225, 2014.

Figure 1a

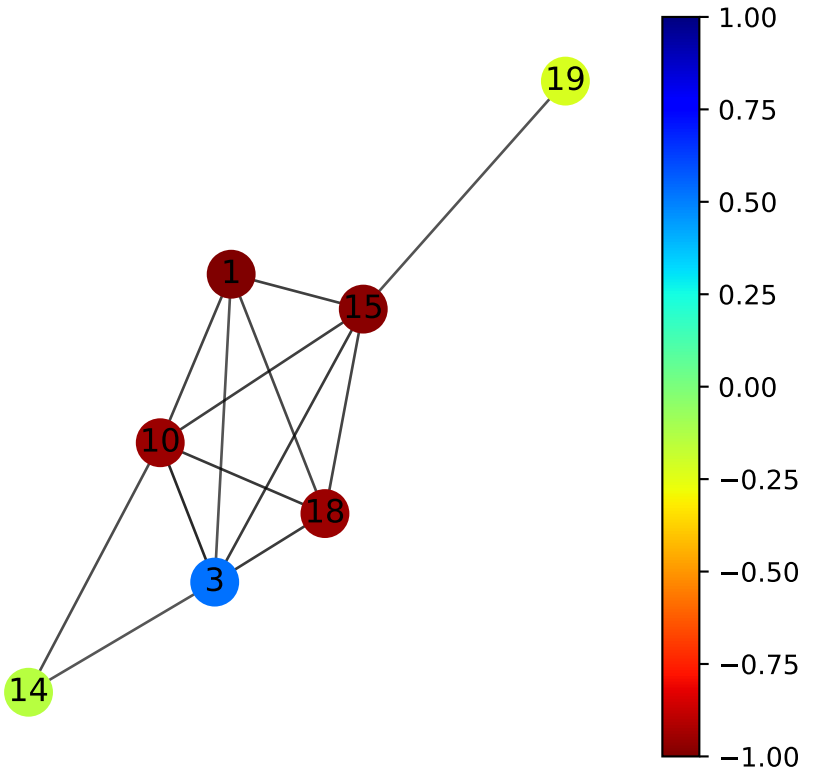

Figure 1b