The Journal of
HISTORICAL
NETWORK
RESEARCH

12 | 2025
special issue

Time and Temporality as Categories of Connectedness

CINDARELLA PETZ | ALINE DEICKE | JULIAN JAROSCH

C²DH
LUXEMBOURG CENTRE FOR
CONTEMPORARY AND DIGITAL HISTORY

uni.lu
UNIVERSITY OF
LUXEMBOURG

The Journal of HISTORICAL NETWORK RESEARCH

ASHLEY DENNIS-HENDERSON/
MATTHEW ROUGHAN/JONATHAN TUKE

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

Journal of Historical Network Research 12 (2025)

Keywords Temporal network analysis, World War I, Natural language processing, Cosine similarity, Semantic similarity networks

Abstract As the number of transcribed and easily accessible historical documents increases, so does the need for computational methods to assist in the analysis of these documents. Over the past two decades, one such computational method that has been applied is network analysis. This paper proposes that temporal network analysis can be applied to historical diaries to find dates and diaries of interest which should be further explored using a traditional close reading approach. More specifically, this paper focuses on how a network of similar diary entries, as determined by cosine similarity, can be created and explored by looking at key metrics such as network density and node centrality. A case study is performed on a collection of Australian World War I diaries held by the State Library of New South Wales to show that the proposed methodology is viable. Furthermore, this paper discusses future work which could improve the potential of this proposed methodology.

C²DH
LUXEMBOURG CENTRE FOR CONTEMPORARY AND DIGITAL HISTORY

UNIVERSITY OF LUXEMBOURG

I. Introduction

As more historic texts are digitized and made easily accessible, there is a growing need to develop and improve methods for computationally analyzing them. This paper aims to illustrate how network analysis can be applied to collections of historical diaries to help inform traditional close reading analysis. Whilst social network analysis has been used for various historical studies, such as Ruth Ahnert and Sebastian Ahnert's *Tudor Networks of Power,*[1] network analysis has not previously been explored as a method for identifying representative documents in a collection.

More specifically, the primary research questions of this paper are:

1) Can network analysis be used to find dates of interest which should be further explored by close reading?
2) Can network analysis be used to find diaries which are representative of a collection?

We propose a methodology to answer these questions based on creating a temporal network of similar diary entries, as determined by cosine similarity scores. To investigate the feasibility of this methodology, we perform a case study using a collection of Australian World War I diaries held by the State Library of New South Wales. We show that this method has potential, with several dates of interest and two representative diaries being found.

The remainder of this paper is laid out as follows. In Section 2 we provide background for the use of computational techniques to analyze humanities data, and provide a review of the relevant literature surrounding the use of network analysis and cosine similarity on historical documents. Section 3 contains a description of the data used in this paper, along with details about how we cleaned this data. In Section 4 we outline our proposed methodology, before providing our results in Section 5. Finally, Section 6 contains a discussion of the outcomes of this paper, along with some proposed future work.

Acknowledgements: We thank the State Library of New South Wales and the AIF Project (University of New South Wales Canberra) for providing the data used in this project. We thank Gareth Pritchard (Department of History, University of Adelaide) for his helpful suggestions regarding the editing of this paper.

Data accessibility: Dennis-Henderson, A., Roughan, M., & Tuke, J. (2025). Supplementary Materials to "Temporal Network Analysis as a Method to Explore Australian World War I Diaries". Zenodo. https://doi.org/10.5281/zenodo.15875366

1 Ahnert and Ahnert 2024.

2. Background

One of the principal pillars of the field of digital humanities is the ability to browse and analyze large quantities of digital texts.[2] To achieve the best results, many researchers suggest that it is beneficial to combine traditional close reading methodology with distant reading techniques, thus utilizing the advantages of both.[3]

Close reading is the careful manual reading and interpreting of a text to identify central themes and their development, allowing a researcher to use their experience and knowledge to help form conclusions and to direct research.[4] Distant reading is the use of computational techniques to create an abstract view of the text and hence visualize global features and trends.[5] These techniques allow large quantities of text to be analyzed, avoiding the need for a researcher to select a small, and possibly biased, sample. Furthermore, distant reading allows researchers to discover and keep track of trends that are only visible across large collections of text.

One such distant reading method, in which there has been a growing interest over the last two decades, is network analysis.[6] However, much of the work in utilizing network analysis to study historical documents has focused on historical social networks.[7] For instance, a recently published book of broad interest is Ruth Ahnert and Sebastian Ahnert's *Tudor Networks of Power*. Within this work, Ahnert and Ahnert investigate a large collection of Tudor letters held in the British government's State Papers to reconstruct and analyze the communication networks at that time.[8] There are many other such works, such as those by Blaine Greteman, Abdurrahman Atçil and Gürzat Kami, and Sam Fields *et al.,* all of which focus on a variety of different documents, historical contexts, and time periods.[9]

Work on historical social networks is important and can provide useful insights into the communities of the past. However, network analysis also has the potential to direct historical research in other ways. In our work we aim to use network analysis not as a method to analyze historical texts, but as a means to locate texts within a collection that should be analyzed further using other methods. As

2 Janicke et al. 2015.
3 Jockers 2013, p. 28.
4 Janicke et al. 2015, p. 84.
5 Janicke et al. 2015, p. 84; Jockers 2013, p. 24.
6 Stein and Fernández Riva 2023.
7 Stein and Fernández Riva 2023, p. iv.
8 Ahnert and Ahnert 2024.
9 Greteman 2021, pp. 2, 13; Atçil and Kami 2022; Fields et al. 2023; for a further discussion on early historical social network analysis see Erickson 1997.

far as we have been able to determine, historical network analysis has not previously been used in this way.

The work that is most similar to ours is that conducted by Henrique de Arruda *et al.* In their work, they propose a method where each node represents *k* sequential paragraphs, with edges between the nodes if they contain similar text, as determined by converting the paragraphs to vector representation using tf-idf (term frequency – inverse document frequency) and calculating the cosine similarity. Using the resultant network, they form a method to identify if a text is real or artificially created and study the development of subjects within a text. However, this work was aimed at the analysis of fiction, such as *Alice's Adventures in Wonderland,* not historical texts.[10]

One of our chief tools, cosine similarity, has been used in many contexts. For instance, Samuel Fernando and Mark Stevenson investigate using cosine similarity to identify paraphrasing within text.[11] Cosine similarity has a long history of use, and we do not attempt a complete survey here, but other uses include text classification, such as the work done by Fawaz Al-Anzi and Dia AbuZeina, and emotion word detection, such as the work performed by Amira El Gohary *et al.*[12] Cosine similarity has the advantage that it is simple to compute and can be applied to texts with unequal lengths. It is also known by other names, including Orchini similarity and the Tucker coefficient of congruence.

3. Data

a. Australian World War I Diaries

The primary data used in this analysis is a collection of 504 Australian World War I diaries held by the State Library of New South Wales.[13] After the armistice in November 1918, Principal Librarian William Ifould and the trustees of the Mitchell Library began the European War Collecting Project.[14] Their aim was to collect high quality, detailed, and descriptive personal diaries of Australian servicemen, regardless of their rank. To achieve this goal, advertisements were placed

10 Arruda et al. 2018; Arruda et al. 2019.
11 Fernando and Stevenson 2009, pp. 1–2.
12 Al-Anzi and AbuZeina 2017, pp. 189, 191; El Gohary et al. 2013, pp. 100, 103–4.
13 "Records of the European War Collecting Project, established by the Trustees of the Mitchell Library in 1918," State Library of New South Wales, https://collection.sl.nsw.gov.au/record/9arGy32n; A copy of the cleaned data used in this analysis is available at: Dennis-Henderson, A., Roughan, M., & Tuke, J. (2025). Supplementary Materials to "Temporal Network Analysis as a Method to Explore Australian World War I Diaries". Zenodo. https://doi.org/10.5281/zenodo.15875366.
14 Edmonds 2020.

in Australian, New Zealand, and British Newspapers in late 1918 and 1919, offering money for suitable diaries. This advertisement was also circulated through the network of the Returned Sailors and Soldiers Imperial League of Australia.[15] Overall, this collection contains over 3,000 individual items, including diaries, letters, narratives, memoirs, and photographs.

The State Library focused on collecting the diaries of Australian servicemen, but their collection also includes diaries from British, French, German, New Zealand, and South African servicemen. Furthermore, due to the close ties between commonwealth countries at that time, not everyone who served in the Australian forces was born (or necessarily lived) in Australia.

As such, for the purposes of this study, we defined our collection of interest to be those diaries written by people with an Australian Imperial Force (AIF) service record that were not nurses or honorary enlistees.[16] Furthermore, we defined diaries as documents which contained a series of dated entries with clear start and end dates.[17] This left us with a collection of 519 diaries written by 199 individuals.

Data cleaning included removing transcribers' notes, HTML code, and any pre- or post-text from the transcriptions.[18] The data was also reformatted from text files containing the complete transcription of each diary into a JSON file for every diary, where each name-value pair contained a single day's entry with its corresponding date. This process was automated using regular expressions to find potential dates and an optimization program to clean these dates.

The date extraction process is reliant on ordered entries, as one would expect in a diary.[19] However, a small number (15) of out-of-order diaries are present in

15 Edmonds 2020, p. 188.

16 This condition excluded the diaries of two nurses, Anne Donnell and Florence Holloway, as well as the official war photographer Frank Hurley, who was given the rank of Honorary Captain. It also excluded 26 diaries written by navy personnel, civilians, or servicemen from other countries, and unit diaries.

17 These documents may have had sections before or after the dated entries detailing things such as letters received, places visited, personal contacts' names and addresses, names of comrades, and pay received. Furthermore, these documents may have had an initial non-dated section where the author wrote about what they had experienced before deciding to start a dated diary. These sections of non-dated text in diaries will be referred to as pre-/post-text.

18 Transcribers' notes are notes, contained within square brackets, that the transcriber wrote to either add additional information, fix spelling or other errors, or to denote words or images that were not or could not be transcribed. The original version of the documents that we acquired contained HTML tags to indicate things such as paragraphs, bold text, underlined text, and crossed-out text.

19 How the accuracy of this date extraction method affects the results will be discussed further in Section 6.

the corpus. These arise, for instance, when the soldier kept dated entries but did not write them in order from the front to the back of the book, seemingly to save paper. Special consideration needs to be given to these diaries. There is an ambiguity of ordering, such that even processing them manually might require us to make arbitrary decisions and thus introduce unnecessary errors into our analysis. Moreover, we are primarily interested in techniques that can be automated to apply to other such sets of historical documents without radical manual intervention. As such, we decided to remove these diaries from this analysis, which gave us a final collection of 504 diaries written by 195 individuals.

Before analyzing these diaries, it is important to understand the scale and structure of the collection. Overall, there are 85,170 entries in this collection ranging from 7th August 1914 to 29th March 1920, as shown in the histogram in Figure 1. This figure shows that more entries were written between early 1916 and late 1918 than during the rest of the war. This is important to note, as this imbalance could affect our results if we do not take it into account. Furthermore, we note that entries continued to be written until early 1920. This was due to the extended time it took to repatriate all Australian soldiers and their dependents back to Australia. Furthermore, many authors continued writing until they were officially discharged.

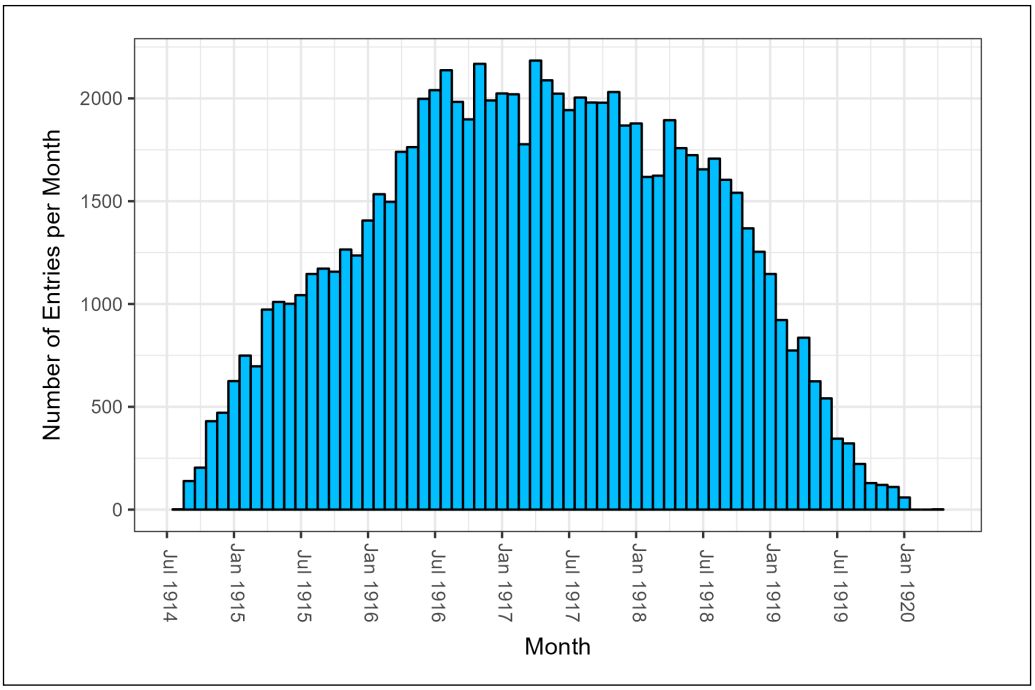

Fig. 1 The number of diary entries in our collection over time. Each bar represents the number of entries in a single month. Note that there are more entries between early 1916 and later 1918 than elsewhere as this was where there was the most overlap in our authors' service periods.

b. Service Records

The diaries used in this study were combined with information about the soldiers obtained from the AIF Project.[20] The AIF project was inspired by a statistical study of the origin and character of the First AIF published by Lloyd Robson in 1973.[21] Alan Gilbert, John Robertson, and Roslyn Russell considered this to be a pioneering study, and still one of the best statistical analyses of the AIF in the late 1980s.[22] However, Robson was only able to sample 0.5% of the records of those who enlisted.[23] As such, in 1987 researchers at the Australian Defence Force Academy (ADFA) began the AIF Project with the aim of creating a comprehensive computer database of AIF service records.[24] This was achieved by combining information from embarkation rolls, the nominal roll, roll of honor circulars, records of the Office of War Graves, and records of the Australian War Memorial.[25]

The information obtained for our authors includes their enlistment date, age, and rank, marital status, and their fate. Figure 2 shows the distribution of enlistment age and date with regards to enlistment rank, and Table 1 gives the breakdown of ranks within our authors. Table 2 provides a breakdown of the marital status of our authors and Table 3 provides a breakdown of our authors' fates when the nominal roll was created in October 1919.

Figure 2 shows that most of our authors were under the age of 35 when they enlisted. This is largely the effect of the maximum permitted enlistment age of the AIF. Originally, the AIF only accepted enlistees between 18 and 38 years old, until June 1915 when the maximum age was changed to 45 years. We note that seven of our authors were aged between 38 and 45 and enlisted after June 1915. Commissioned officers were not subject to this age restriction, and we observe that four of our authors who were commissioned officers enlisted early in the war, all of whom were over the age of 50. Closer inspection of the data shows that three of these authors were doctors, and one had previous experience in the Boer War. The histogram at the top of Figure 2 shows that there are two peaks in the enlistment dates of our authors, one in late 1914 when the war began and the other in late 1915. Aside from these peaks, the enlistments of our authors continually decreased over time, with only two of our authors enlisting after March

20 "The AIF Project," UNSW Canberra, https://www.aif.adfa.edu.au/index.html; the data gained from the AIF Project is the intellectual property of and copyrighted by UNSW Canberra and was provided to us for this specific research project. As part of our agreement with UNSW Canberra we are unable to provide a copy of this data.
21 Gilbert et al. 1989, p. 107.
22 Gilbert et al. 1989, p. 107.
23 Gilbert et al. 1989, p. 107; Robson 1973, p. 738.
24 "The AIF Project" 1993, p. 34; Gilbert et al. 1989, p. 108.
25 "The AIF Project" 1993, pp. 34–6.

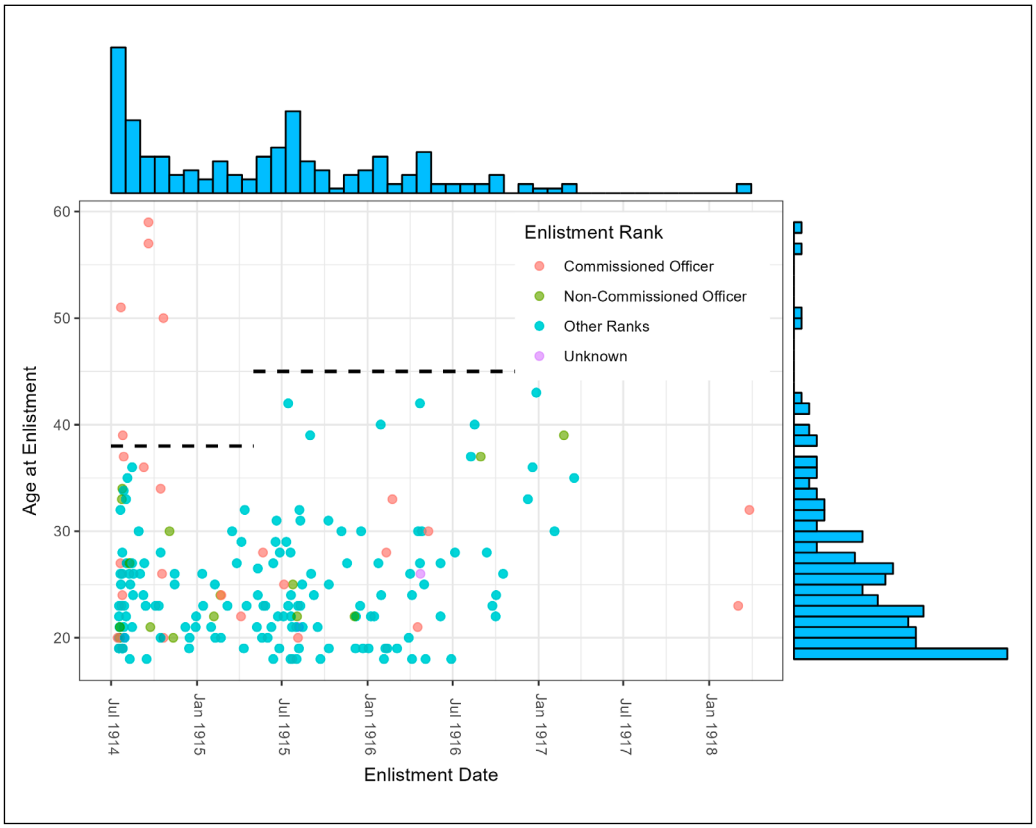

Fig. 2 Scatterplot showing the enlistment age, date, and rank for our authors alongside histograms of their enlistment age and dates. Four authors with unknown enlistment age/date are not shown here. The black dashed lines represent the maximum enlistment age for the AIF. This was originally 38 years, until June 1915 when it changed to 45 years. This age limit was not applied to commissioned officers, as can be seen on the plot. We observe that seven of the authors were between 38 and 45 years old and enlisted after the age changed in June 1915. All but two of the authors, Ormund Tunks and Edy Little, enlisted prior to May 1917.

	Commissioned Officer	Non-Commissioned Officer	Other Ranks	Unknown
Count	26	17	148	4
Percentage	13.3%	8.7%	75.9%	2.1%

Tab. 1 The breakdown of enlistment ranks (commissioned officers, non-commissioned officers, other ranks, or unknown) amongst our World War I authors.

	Married	Single	Widower	Unknown
Count	29	164	1	1
Percentage	14.9%	84.1%	0.5%	0.5%

Tab. 2 The breakdown of marital status (married, single, widowed, or unknown) amongst our World War I authors. Most of the authors were single.

	Returned	Died	Abroad	Discharged	Unknown
Count	164	14	5	2	10
Percentage	84.1%	7.2%	2.6%	1%	5.1%

Tab. 3 The breakdown of fate (returned to Australia, died in World War I, effective abroad, discharged, or unknown) amongst our World War I authors when the nominal roll was created in October 1919. Most of the authors had returned to Australia by that time.

1917. This follows the general trend of enlistments in the AIF, which were initially high and decreased over time.

From Table 2 it can be observed that most of our authors were single. This is unsurprising as it would have been more difficult to enlist with dependent family members.

Table 3 shows that by October 1919 most of our authors had returned to Australia, whilst five authors were still abroad, and two had been discharged. However, we note that only 7.2% of our authors had died during the war, whereas approximately 18% of Australians who embarked in World War I died.[26] This means our collection is biased towards those who survived the war, and hence were possibly in safer roles. Although this does introduce some bias into the data, this bias is inevitable as we rely on diaries which are often lost unless the author survives. It is important to be aware of this during our analysis. However, this inherent bias in our data will not detract from the usefulness of our methodology.

26 Embarkation and death figures vary by source. The Australian War Memorial notes 331,781 AIF embarkations and 60,284 deaths giving a figure of 18.17%. Bill Gammage, in his book *The Broken Years,* notes three different sets of statistics: 330,700 embarkations and 63,163 deaths (19.1%), 331,781 embarkations and 59,330 deaths (17.88%), and 331,814 embarkations and 58,132 deaths (17.52%); "Enlistment statistics, First World War," Australian War Memorial, last modified March 18, 2022, https://www.awm.gov.au/articles/encyclopedia/enlistment/ww1; Gammage 1990, p. 313.

4. Methodology

a. Creating the Temporal Network

Initially, a temporal network is created where the nodes are individual diaries. On each date, two nodes are connected by an edge if both diaries have entries for that date and those entries have high similarity, with the edge weight being the similarity score. Similarity scores are determined by first converting the entries to a vector representation using BERT and then calculating the cosine similarity between these vectors.[27]

Cosine similarity is a measure of the similarity between two vectors, based on the angle between the vectors. Suppose we have two vectors, $\boldsymbol{a}$ and $\boldsymbol{b}$, then the cosine similarity is given as

$$sim(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a} \bullet \boldsymbol{b}}{\|\boldsymbol{a}\| \|\boldsymbol{b}\|},$$

where $||\boldsymbol{a}||$ is the length (or Euclidean norm) of the vector $\boldsymbol{a}$, and $\boldsymbol{a} \bullet \boldsymbol{b}$ is the dot product.[28] Cosine similarity scores range from –1 to 1. A similarity score of 1 represents identical texts, and a similarity score of –1 represents texts that are completely different. Therefore, as the cosine similarity score gets closer to 1, the entries become more similar.

There are several other vectorization methods and similarity measures for text (e.g. tf-idf and Jaccard similarity). However, BERT and cosine similarity were chosen due to their extensive use in the digital humanities literature as well as their simplicity, both in understanding their output and in their implementation. This simplicity makes it easier for researchers in a variety of fields to utilize our methods. This choice will be further discussed in Section 6.

Although a cosine similarity score is calculated for every pair of entries with the same date, we cannot create a network edge between every pair as the 'network' would be trivial. As such, edges are only formed when it is deemed that the similarity between a pair of entries is high. However, the term 'high similarity' is subjective, and there is no standard, agreed upon method for determining what constitutes high similarity. For instance, Rada Mihalcea, Courtney Corley, and Carlo Strapparava use 0.5 as their similarity threshold when considering the semantic similarity of short text snippets.[29] Ulfa Mardatillah *et al.,* whilst studying the similarity between scientific articles and the articles they cite, define 0–0.3 as

27 Devlin et al. 2019; Han et al. 2011, pp. 77–8.
28 Han et al. 2011, pp. 77–8.
29 Mihalcea et al. 2006, pp. 775, 778.

low similarity, 0.3–0.6 as medium similarity, and 0.6–1 as high similarity.[30] However, neither of these studies present reasoning as to why they have chosen these values.

To add to this confusion, the authors de Arruda *et al.* note that it is impractical to choose a single constant threshold value as every dataset is different, and in some cases the threshold could remove all or none of the edges. As such, they considered a set of thresholds that kept the top k% of similarity scores, and then investigated which thresholds would give the network properties they required.[31]

Perhaps the best method for determining a threshold value might be to take a selection of entries and manually judge which are highly similar. This data could then be used to determine an appropriate threshold. However, this is *very* time consuming, and therefore not practical or realistic for many historians if they wished to apply this methodology to help inform their analysis. As such, we employed a method similar to de Arruda *et al.* We chose a similarity threshold such that the top k% of similarity scores were kept, with the value of k determined by performing a sensitivity analysis. That is, we compared the results of our methods for each threshold and used the results to determine if, and how, the threshold affects the results, seeking a robust choice.

b. Finding Dates of Interest

Our first research question considers whether we can use network analysis to identify dates of interest from a collection of historic diaries. To do this, we consider the static network for each date which is created by our temporal network. The sentiment of each diary entry, calculated using VADER,[32] was added to these networks as a node attribute.[33] Three metrics were then considered for each date and associated network to determine which days are of interest.

A date of interest would be one where many authors are discussing the same thing, signifying that something interesting or important happened that day. As such the first metric considered is the number of nodes on each date. However, it is not enough to simply have a date with many nodes – these nodes must also be highly connected, showing that the authors were all discussing the same thing. As such, the other two metrics considered are the density and transitivity of each date.

30 Mardatillah et al. 2021, p. 3.
31 De Arruda et al. 2018, pp. 129, 131.
32 We consider the suitability of VADER in this scenario further in Section 6.
33 Hutto and Gilbert 2014.

The density of a network is the ratio of the number of edges that exist to the number of edges that could exist. Mathematically, for an undirected graph, this can be written as

$$density = \frac{2m}{n(n-1)},$$

where *n* is the number of nodes and *m* is the number of edges in the network. The density of a network takes a value between 0 and 1, where density 0 represents a network with no edges and density 1 represents a network with all possible edges. For our static networks, the density can be thought of as the probability that any two randomly chosen entries would have high similarity.[34]

The transitivity of a network is the ratio of the number of triangles that exist to the number of connected triples that exist. Mathematically, this is written as

$$transitivity = \frac{(number\ of\ triangles) \times 3}{(number\ of\ connected\ triples)}$$

Here, a triangle is a set of three nodes with edges between each pair of nodes. That is, if we have three nodes *u*, *v*, and *w*, then there are edges (*u*, *v*), (*u*, *w*), and (*w*, *u*). A connected triple is a set of three nodes with at least two of the edges present. That is, there are edges (*u*, *v*) and (*v*, *w*), while the edge (*w*, *u*) may or may not exist. The factor of three in the previous equation accounts for the fact that each triangle has three sets of connected triples.[35]

For this study, when considering dates with high transitivity it is important that they also have high density. Consider the diagram shown in Figure 3. The network in Figure 3-A has high transitivity but low density. Therefore, we have two cliques which are writing about separate subjects. The network in Figure 3-B has both high transitivity and high density, and so everyone is writing about the same subject.

c. Finding Diaries of Interest

Our second research question considers whether we can use network analysis to identify diaries that are representative of the entire collection. To investigate this, our temporal network is converted into a single static network where each node is a diary, edges exist between two nodes if those diaries have similar entries on at least one day, and the edge weights are the number of days in which those diaries have similar entries. We then consider two metrics for each node to determine diaries of interest.

34 Newman 2018, p. 128.
35 Newman 2018, pp. 183–5.

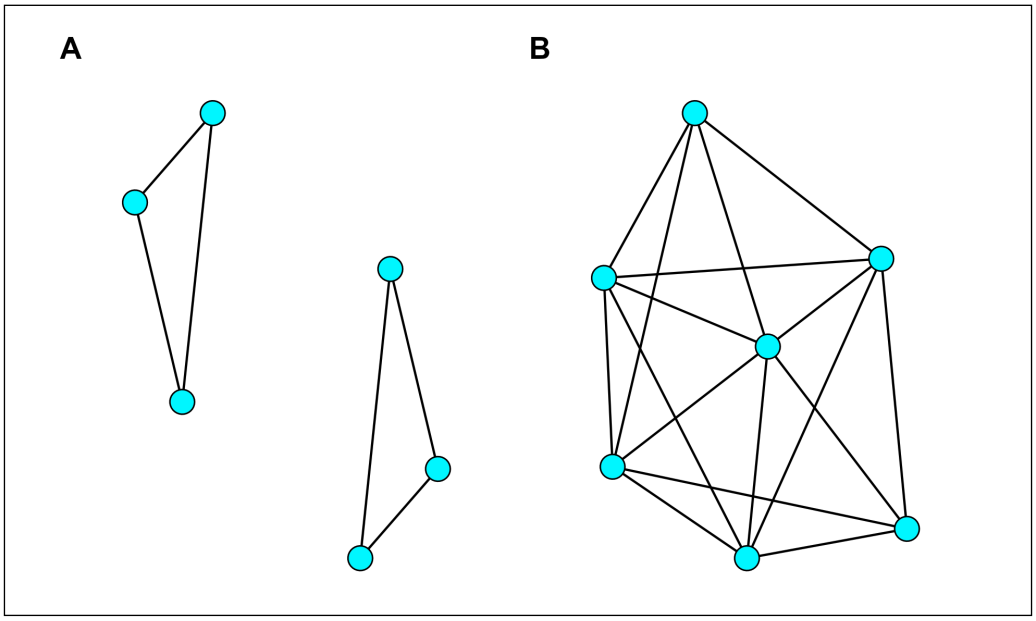

Fig. 3 (A) A network with transitivity of 1 and density of 0.43. (B) A network with transitivity of 0.76 and density of 0.81.

Diaries of interest are those that have the most in common with all other diaries. Hence, our metrics focus on node centrality. The two metrics we consider are weighted node degree and weighted eigenvector centrality.

The degree of a node is the number of edges connected to it.[36] Hence, in terms of our network, the node degree is the number of diaries which have a similar entry to the diary of interest on at least one date. This definition can be expanded to include the edge weights, giving us the weighted node degree as the sum of the edge weights for the edges connected to the node. In terms of our network, the weighted node degree would tell us how many times our diary of interest has similar entries (on the same date) to other diaries. Therefore, when looking for a diary of interest we want a diary with high weighted node degree.

An extension of node degree is eigenvector centrality. When considering node degree, all neighbors of a node are considered to have equal importance. However, in eigenvector centrality we consider whether a node's neighbors are also important in the network. In a social network, node degree would be equivalent to how many people you know, whereas eigenvector centrality would also consider who you know. In terms of our network, if our diary of interest is similar to another diary which is highly similar to the rest of our collection, then our diary

36 Newman 2018, p. 126.

of interest is also highly similar to the rest of our collection. For an undirected network, the eigenvector centrality of node *i* is the i^{th} element of the eigenvector for our graph's adjacency matrix corresponding to the largest eigenvalue.[37]

5. Results

a. Determining the Similarity Threshold

After calculating a similarity score for every pair of non-empty entries that are on the same date, we have 2,241,590 similarity scores ranging from –0.345 to 1. A histogram with overlaid density curve of these scores is shown in Figure 4.

However, as previously noted, retaining all the edges in our network is impractical. As such, it is necessary to determine a similarity threshold, keeping only edges with a similarity score above the threshold.

A range of thresholds were tested and the metrics for both research questions were considered. The thresholds we investigated were the points where the top *k*% of all the edges were kept, with *k* ranging from 5 to 50, in increments of 5. The similarity scores at these thresholds are given in Table 4. Figure 5 shows the metrics for our first research question at various thresholds, whilst Table 5 and Table 6 show the top two diaries in terms of the metrics for our second research question at different thresholds.

From Figure 5, it is noted that the metrics follow the same pattern, except at the extremes of Figure 5-A due to the lack of data for those extremes (as noted in Figure 1). Of particular interest is that as the threshold value increases the density curve flattens, and as the threshold value decreases the transitivity curve flattens. As the pattern is the same, any of these thresholds should produce similar results. However, we want a threshold with information for both transitivity and density, so we chose 0.613, which is in the middle. This threshold keeps the top 25% of the similarity scores.

Table 5 shows that when we sort our diaries by weighted eigenvector centrality and then weighted node degree, the top two diaries remain the same for threshold values between 0.487 and 0.613, and then begin to change. Table 6 shows that when we sort our diaries by weighted node degree and then weighted eigenvector centrality, the top two diaries remain the same for all threshold values except 0.736. Due to this consistency, any threshold we choose below 0.613 will give us the same results for the top two diaries of interest. As such, 0.613 appears to be, once again, suitable as our threshold.

37 Newman 2018, pp. 159–61.

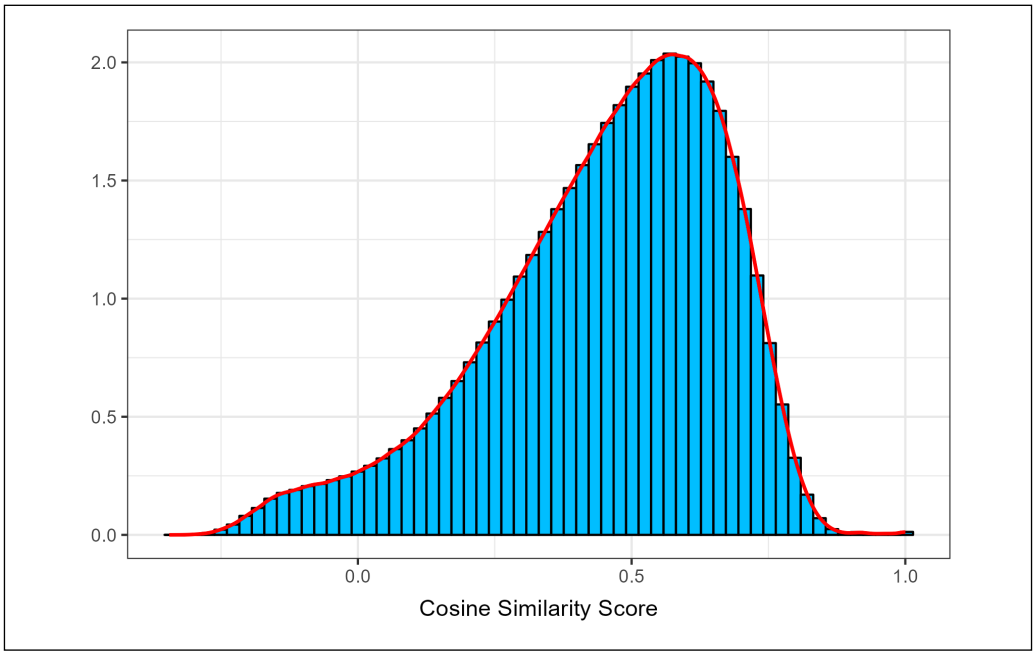

Fig. 4 Histogram with overlaid density curve of the cosine similarity scores.

Percentage	50%	45%	40%	35%	30%	25%	20%	15%	10%	5%
Similarity Score	0.487	0.514	0.539	0.564	0.588	0.613	0.639	0.666	0.696	0.736

Tab. 4 Table of percentages and corresponding similarity scores.

Threshold	0.487	0.514	0.539	0.564	0.588	0.613	0.639	0.666	0.696	0.736
Diary 1	401	401	401	401	401	401	401	72	72	72
Diary 2	355	355	355	355	355	355	72	311	311	97

Tab. 5 The top two diaries in terms of weighted eigenvector centrality, followed by weighted node degree.

Threshold	0.487	0.514	0.539	0.564	0.588	0.613	0.639	0.666	0.696	0.736
Diary 1	401	401	401	355	355	355	355	355	355	355
Diary 2	355	355	355	401	401	401	401	401	401	72

Tab. 6 The top two diaries in terms of weighted node degree, followed by weighted eigenvector centrality.

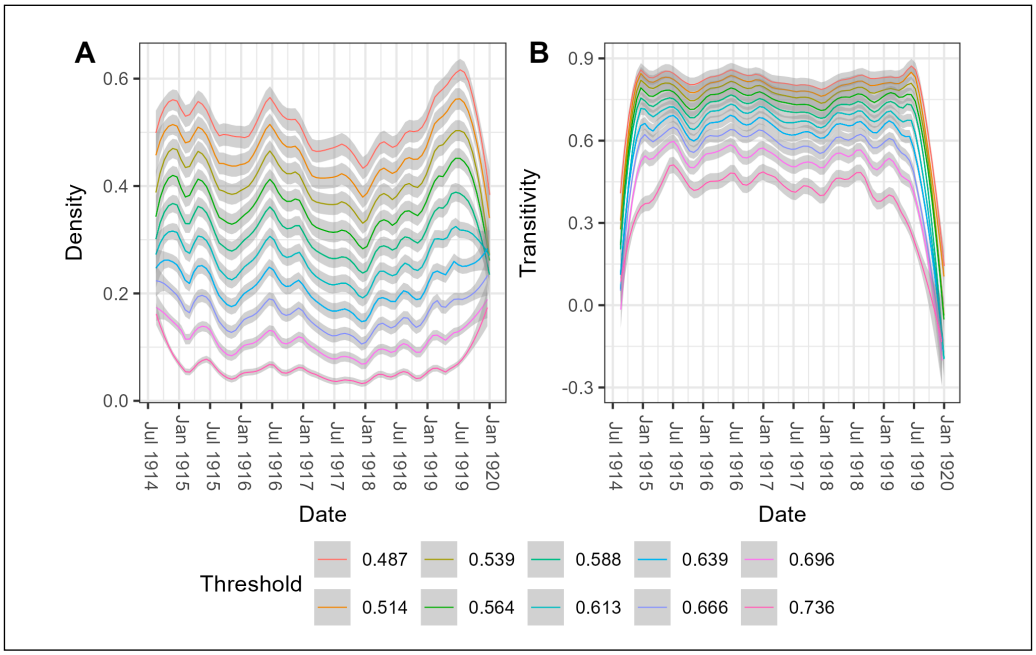

Fig. 5 (A) Network density for each date and a variety of thresholds. (B) Network transitivity for each date and a variety of thresholds. Smoothing has been applied to both graphs to improve readability. It is noted that aside from the extremes of graph A, density and transitivity both follow the same pattern over time for all threshold values. However, as the threshold increases, the density curve flattens, whilst as the threshold decreases the transitivity curve flattens.

b. Finding Dates of Interest

As previously mentioned, the dates of interest will be those where many people are discussing the same topic. As such, we consider the metrics of number of nodes, density, and transitivity for each date. As our metrics do not have simple linear relationships, we cannot simply find dates with the highest values in all three metrics. Instead, we plot these metrics, as shown in Figure 6, and seek outliers.

Six dates of interest have been highlighted in Figure 6. These dates have been chosen as they follow along the top edge of the curve but are separate from the bulk of the distribution. All these dates have higher density than other dates with a similar number of nodes. Furthermore, they all have high transitivity. It is to be noted that there are three dates occupying the same point in the top left-hand corner of the graph which have not been chosen despite having high density and high transitivity. This is because those dates only have three nodes, and hence do not meet our criteria of dates where *many* people are talking about the same thing. The six dates selected are all important dates. Those highlighted are

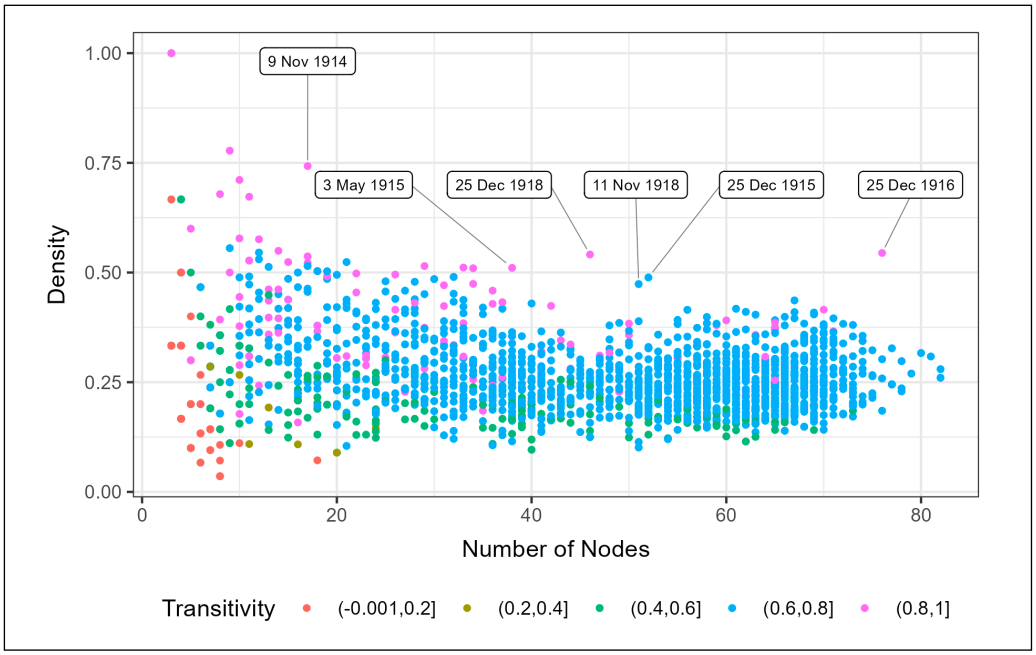

Fig. 6 Scatterplot of the number of nodes and density for the static network of each date, highlighted by the network's transitivity. The labelled dates of interest are those along the top of the curve, but somewhat separate from the bulk of the data.

9 November 1914 (Battle of the Cocos), 3 May 1915 (Gallipoli Campaign), 11 November 1918 (the Armistice), and Christmas 1915, 1916, and 1918.

Most of these dates are obvious selections for any war historian, which is confirmation that the technique is working correctly. However, the Battle of the Cocos may be less familiar to those outside Australia. Even knowing of the battle, without knowing that some diary authors were present at this smaller engagement, it might not be a date one would arbitrarily choose to examine closely.

On 9 November 1914 the first Australian convoy was passing the Cocos Islands when they received a mayday signal from the Cocos wireless station stating that a strange warship was approaching. The *HMAS Sydney,* an Australian light cruiser protecting the convoy, responded to the signal, spotting and engaging the German light cruiser *SMS Emden,* which had just destroyed the cable station on Direction Island.[38] The *Sydney* was able to defeat the *Emden,* forcing the *Emden*'s commanding officer to beach her to prevent sinking.[39]

38 Bean 1941, pp. 103–7.
39 Olson 2018.

Seventeen authors in our collection wrote diary entries on 9 November 1914, and their similarity network is shown in Figure 7. All but two of these entries discuss the Battle of the Cocos. Whilst the entries differ in presentation, these authors were all located on ships in the convoy that *Sydney* was protecting. The authors discuss the *Sydney*'s chase and defeat of the *Emden* and its collier, the number of wounded and dead from the *Sydney,* and the celebrations that happened onboard upon hearing the news. Many of them also comment on the weather (warm, yet rainy), illness and death within the convoy, and travelling with no lights at night. The differences in sentiment seen in Figure 7 stem from how the authors reported what happened. Those with positive sentiment focused more on how lucky the *Sydney* got and the ensuing celebrations within the convoy. Those with negative sentiment focused more on the sick, wounded, and dead. Those with neutral sentiment either commented on the positives and negatives in equal proportion or used very matter-of-fact language when writing their entry for this date. The two entries that do not discuss this battle are Walter Rainsford's, whose entry is simply "Nursing Practice", and Clarence Read's, who was not in the convoy but stationed in Papua New Guinea. Rainsford's entry does not appear in Figure 7 as it was not deemed similar to any others, but Read's does as that day he was chasing boats around Papua New Guinea and salvaging wreckage.

Thirty-eight authors in our collection wrote diary entries on 3 May 1915, and their similarity network is shown in Figure 8. Of these authors, five do not have highly similar entries to any others, and hence are not shown. The remaining thirty-three authors fall into two groups: those discussing the fighting at Gallipoli and the Dardanelles, and those training in Egypt. Interestingly, from just looking at the graph it is not possible to distinguish these groups, despite one of them being an active war zone, whilst the other is not. This is likely because both groups discussed the wounded – those at Gallipoli from witnessing the wounding, and those in Egypt from seeing wounded brought back to hospitals there. The other interesting thing about this date is that while it was part of the Gallipoli Campaign, 2–3 May 1915 was specifically the Battle for Baby 700, one of the hills. However, none of the authors mention Baby 700. The closest mentions are by James Baillie, who discusses starting at Pope's Hill and advancing to the next ridge, and Elis Silas, who mentions storming a ridge. Baillie, and likely Silas, were taking part in the battle for Baby 700, but do not refer to it as such. This is in stark contrast to the diary entries from the Battles of the Cocos where all the authors seemed to parrot the same information. This highlights the effects of being on the front line, and not necessarily knowing the specifics of what you are doing or what is happening around you.

It is unsurprising that Christmas and the Armistice appear here as they would have been days of commonality for all soldiers, no matter where they were. What is of interest is that on the networks for these dates, despite being generally considered as positive events, some of the entries are negative. Looking more in depth at Christmas 1916 (Figure 9-B) it is noted that many of the negative en-

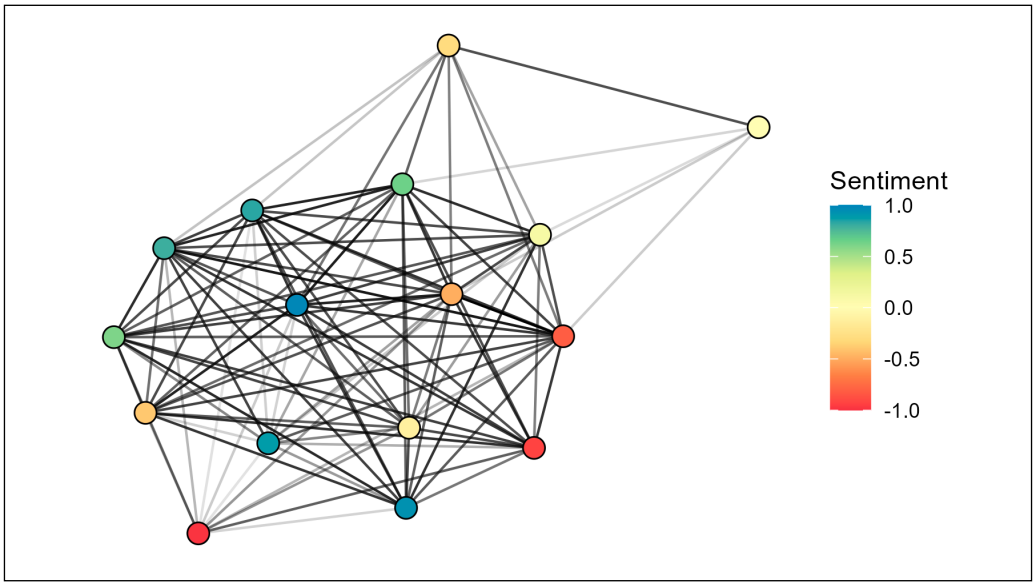

Fig. 7 Network for 9 November 1914. Each node represents an entry from a single diary, colored based on its sentiment which ranges from –1 (very negative) to 1 (very positive). Seventeen authors wrote entries on this date. Fifteen were discussing the Battle of the Cocos, one was discussing his work around ships in Papua New Guinea, and one (not shown) was commenting on nursing practice.

tries were from soldiers on the front lines who spent a cold and muddy Christmas under enemy fire with a lack of good food.

Another way to approach the identification of interesting dates is to look at the networks in the top *i*% of the metric values. For example, Table 7 shows dates for the top 10 % of density and transitivity values and the top 50 % of number of nodes. Note that most of the dates are in 1916. This is due to the imbalance in our data, as seen in Figure 1. Therefore, it is necessary to be careful when looking at dates with the highest number of nodes.

This method has been able to highlight interesting dates for close reading analysis. Importantly, although some of these dates may have been expected, i.e., the Armistice, others were not. For instance, prior to this analysis we were aware that no-one in our collection was on the *Sydney* and therefore did not expect for the Battle of the Cocos to be highlighted.

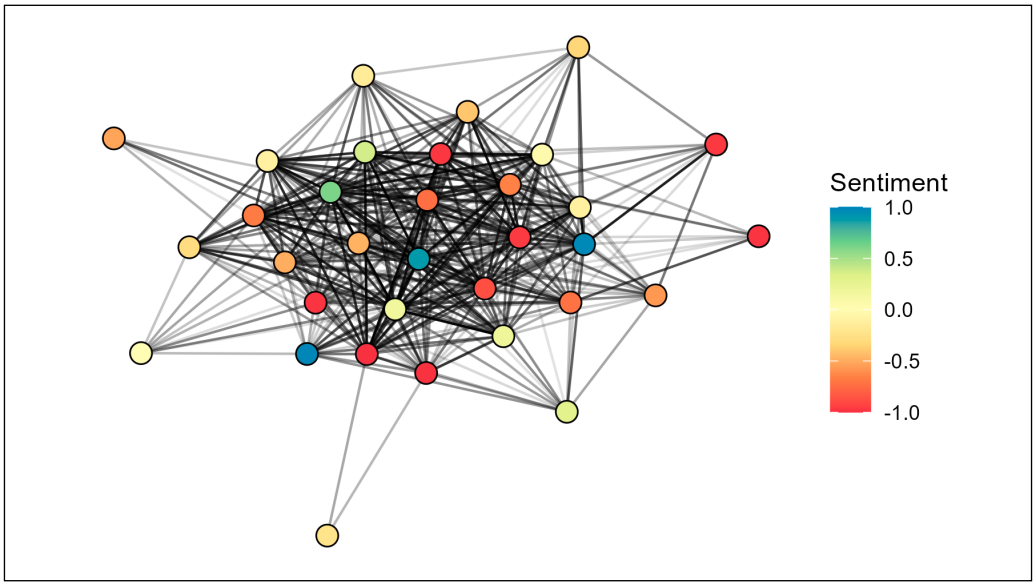

Fig. 8 Network for 3 May 1915. Each node represents an entry from a single diary, colored based on its sentiment which ranges from –1 (very negative) to 1 (very positive). Thirty-eight authors wrote entries on this date, although only thirty-three of them have high similarity with other entries.

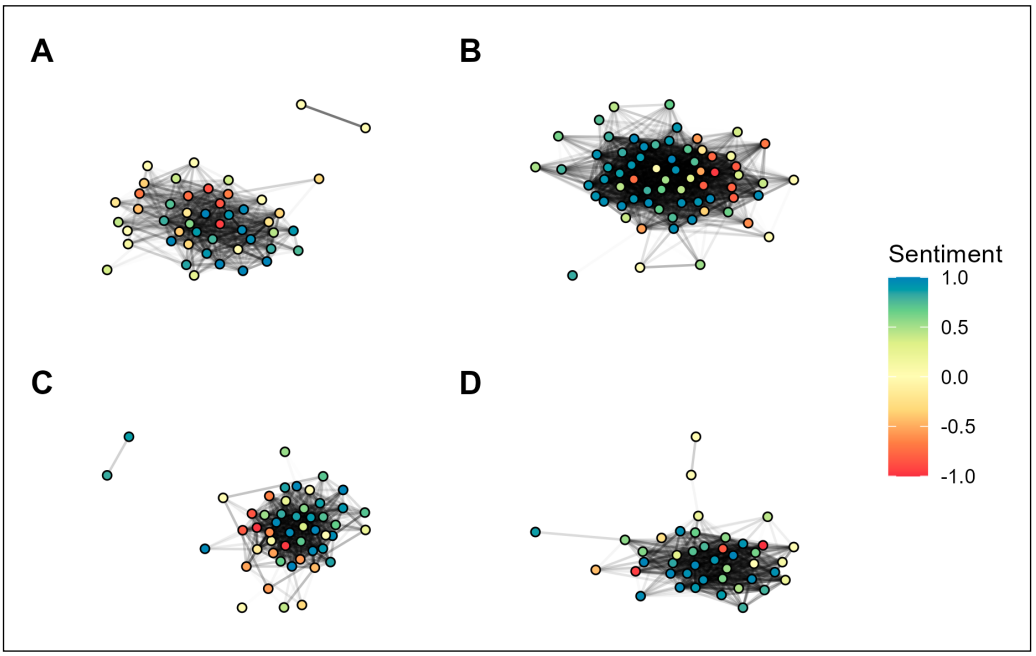

Fig. 9 Networks for (A) 25 December 1915, (B) 25 December 1916, (C) 11 November 1918, and (D) 25 December 1918. Each node represents an entry from a single diary, colored based on its sentiment which ranges from –1 (very negative) to 1 (very positive).

Date	Number of Nodes	Density	Transitivity
25 Dec 1915	52	0.489	0.776
26 Jan 1916	51	0.389	0.775
24 Apr 1916	58	0.399	0.796
13 May 1916	60	0.391	0.826
11 Jun 1916	67	0.396	0.785
22 Jul 1916	65	0.391	0.791
20 Aug 1916	65	0.374	0.835
31 Oct 1916	65	0.386	0.807
24 Dec 1916	70	0.377	0.799
25 Dec 1916	76	0.545	0.838
25 Dec 1917	70	0.415	0.839
25 Mar 1918	59	0.380	0.810
30 Mar 1918	58	0.370	0.800
28 Sep 1918	50	0.384	0.832
11 Nov 1918	51	0.474	0.776

Tab. 7 List of dates in top 10 % of density and transitivity scores and top 50 % of number of nodes.

c. Finding Diaries of Interest

The diaries of interest in this collection will be those that have the most in common with all the other diaries, and hence have a higher centrality value. The two centrality measures we consider are weighted node degree and weighted eigenvector centrality. To find our diaries of interest we again graph the metrics, as shown in Figure 10, and look for outliers. Diaries that have high values for both metrics are 355 and 401.

Diary 355 was written by William Osborne, a driver mechanic from Queensland. Osborne was originally born in England but was living in Australia when he enlisted on 29 Feb 1916 at age 24 with the rank of Motor Transport Driver. He embarked from Melbourne on the Medic on 16 Dec 1916 and returned to Australia in August 1919. His diary covers the period between his embarkation and returning to Queensland on 22 August 1919.

Osborne's diary begins by discussing life on the troop ship after leaving Australia. On board, they occupy their time with lectures, drills, parades, concerts,

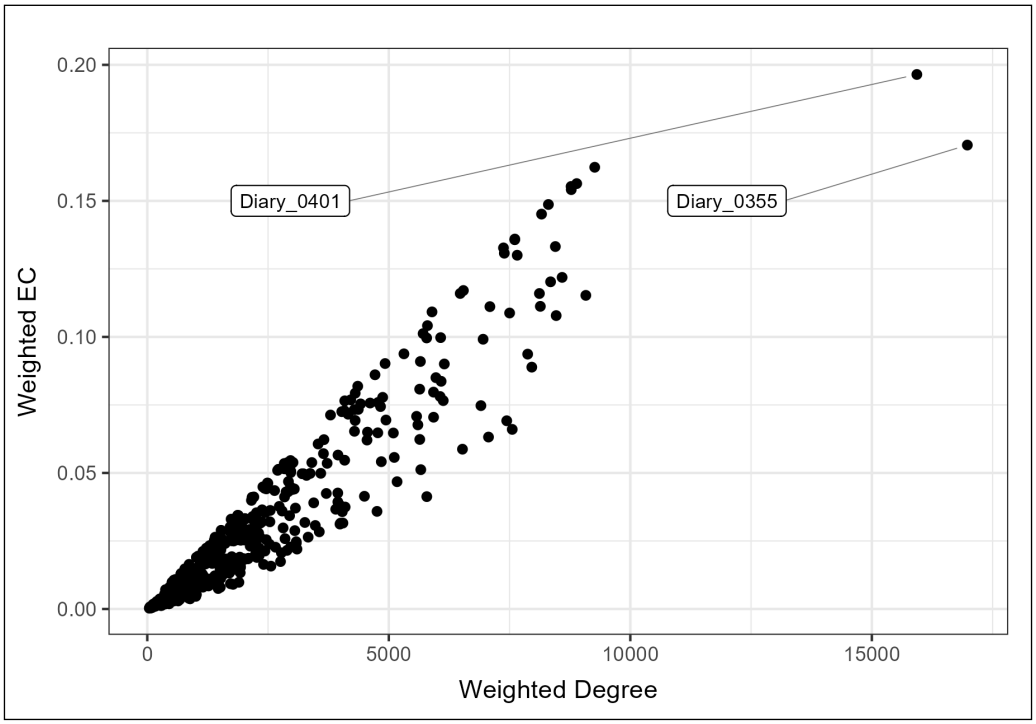

Fig. 10 Weighted degree and weighted eigenvector centrality for all our diaries.

sports, card games, reading books, and writing letters. Furthermore, across this journey Osborne discusses sickness and death on board, the weather, sailing in convoy, the threat of submarines, and the equator-crossing ceremony featuring King Neptune. After arriving in England, Osborne discusses their journey to camp, and various activities they did whilst there, including being issued supplies, getting inoculations, undertaking medical examinations, marching, parading, drilling, guard duty, escorting prisoners, being paid, writing letters, visiting the local town, and tobogganing. After crossing the channel to France, Osborne began working with vehicles, in particular lorries. His duties included maintaining and repairing the vehicles, as well as moving supplies and troops to where they were needed. Whilst Osborne did not fight in any battles, he often performed his duties under fire, and occasionally had to evacuate civilians or wounded soldiers. Furthermore, in his off time he saw many locations that had previously been bombed or had fighting. Aside from performing his duties, Osborne was also able to take leave and visit England, Scotland, and Paris. On 11 November 1918 Osborne related the excitement caused by the Armistice, and his day-to-day duties began to change. Initially, he still worked on maintaining vehicles and transporting supplies, but he no longer worked under the threat of bombs and had more time to visit local towns around France and Belgium. As the first step in being demobilized, Osborne was sent to England, where he again had free time to sightsee and visit family. On the journey back to Australia he undertook many of the same

pastimes as the journey over. After arriving back in Queensland, he was given a leave pass whilst pending discharge.

Osborne's diary gives the reader an overview of the life of a transport worker during World War I. His diary would be like others who worked in roles close to, but not on, the front line. It gives us an idea of the journeys to and from Australia, as well as leave and medical treatment for non-officers.

Diary 401 is the first of two diaries written by Sir Charles Rosenthal. Prior to World War I, Rosenthal was an accomplished architect and singer. He also had an interest in militia soldiering, and in 1903 had received a commission as a second lieutenant in the Militia Garrison Artillery. In 1908 he transferred to the Australian Field Artillery, and in 1914 he became the commanding officer of the 5th Field Artillery Brigade. On 25 August 1914, Rosenthal joined the AIF as the commander of the 3rd Field Artillery Brigade with the rank of Lieutenant-Colonel. He embarked with the first convoy on 25 September 1914 and was still serving abroad when the nominal roll was created in October 1919.[40]

Rosenthal's diary is a very detailed account of his activities throughout the war. Rosenthal's journey to Egypt was very similar to the account given by Osborne, with the addition of having responsibilities to ensure smooth passage for the men and horses. Whilst in Egypt, Rosenthal spent much of his time organizing camp, giving lectures, and meeting with other officers. However, he still found time to go sightseeing, write letters to loved ones, and attend concerts. During the Gallipoli Campaign in 1915 Rosenthal was often near the front line, doing reconnaissance and organizing troops and guns. During this time, he was twice evacuated from Gallipoli, once due to injury and once due to illness. From June 1916 onwards Rosenthal oversaw various units on the Western Front and took part in several important battles. He spent his days travelling large distances to inspect and organize troops, reconnoiter the land, and visit the wounded. He also took part in capturing prisoners. During this time, he was wounded again and gassed.

Rosenthal's diary gives the reader insight into the experience of the front lines as well as the organization of troops. Furthermore, this diary covers many common aspects of a soldier's career, including injury and illness, leave, communication with the home front, and transport. Finally, Rosenthal's diary provides insight into the roles of officers during the war.

40 Dennis et al. 2009; Hill 2006.

Overall, the two diaries highlighted by this method cover many aspects of the Australian experience in World War I, including

- travelling overseas,
- training in both Egypt and England
- battles in both Turkey and Europe
- the differences between the experiences of enlisted and commissioned officers
- illness and injury
- communication with the home front

Furthermore, both diaries cover a significant period of the war. Based on a previous manual examination of this collection, we are also aware that the topics discussed in these two diaries cover the main topics within the entire corpus. Thus, our approach has identified two representative diaries for this collection.

6. Discussion

The results of this study show that the proposed methodology is feasible and can produce useful results. However, there are several decisions that were made in the workflow, for example the choice of a cosine similarity threshold, that warrant further research to determine the optimal choices.

To calculate the similarity scores between entries, this study used BERT to find a vector representation and cosine similarity to calculate the score. BERT was chosen as it has become a well-known and easy to implement standard approach.[41] Likewise, cosine similarity was chosen for the same reason. However, there are various other methods for determining the similarity between two texts. For example, tf-idf or Word2Vec could be used instead of BERT, with Jaccard similarity. Whilst this study has shown that the combination of BERT and cosine similarity provides useful results, in the future it would be ideal to apply a range of different methods for creating vector representations and calculating similarity to determine if this significantly affects the results of the analysis. Likewise, the method for choosing a similarity threshold should be examined further.

Due to the various ways in which dates were written in these diaries, as well as potential errors by both author and transcriber, it is not possible for our automatic date extraction method to be 100 % accurate. We manually extracted the dates from ten diaries and compared these to the automatically extracted dates we found. For each date, we calculated the distance between the manually and automatically extracted dates, and a bar graph of these distances is shown in Fig-

41 Mickus et al. 2020, p. 350.

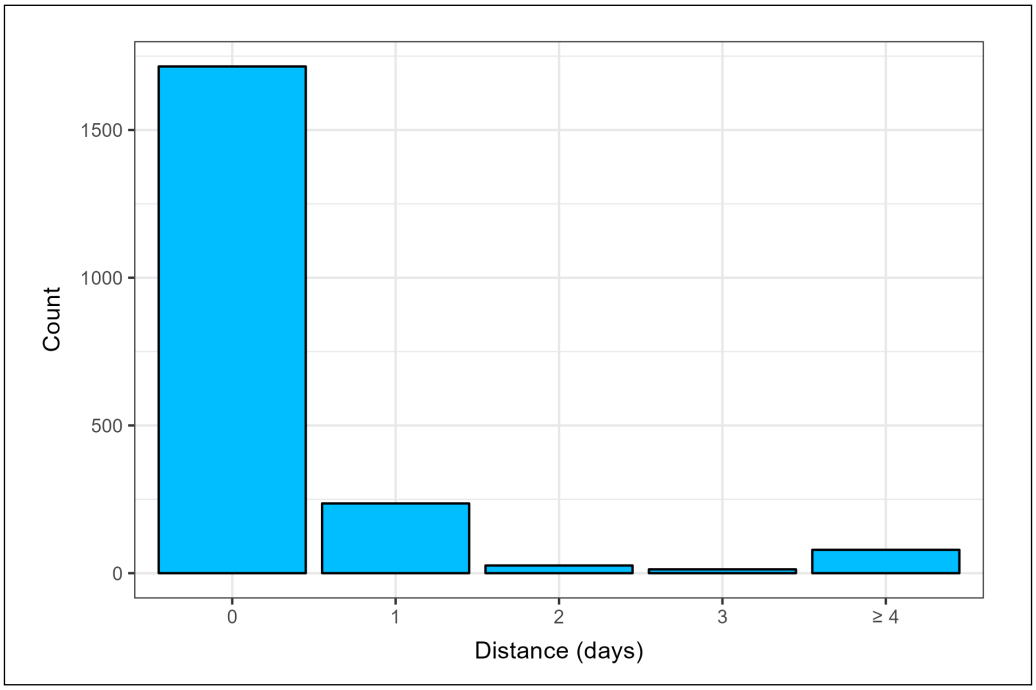

Fig. 11 Bar graph of the distance between our manually and automatically extracted dates. Most of the dates are within three days.

ure 11. From this we note that the method was completely accurate for a large proportion of the dates, and that most dates were within three days of what they should have been.

The lack of complete accuracy in the automatic date extraction would have impacted the results of the network analysis. When creating our temporal network, we calculate the similarity between entries on the same date, under the assumption that entries written on a particular day detailed the events that happened that day. However, inaccuracies in the automatically extracted dates would mean we sometimes calculate similarity between entries that were written on different dates.

One way to overcome this is to ensure that all the extracted dates we have are completely accurate. However, this would not be possible with an automatic extraction method. As such, the dates would have to be manually extracted, which would take a large amount of time. Furthermore, this may still not be completely accurate due to human error.

Rather than striving for completely accurate dates, the other possible way to overcome this issue is to consider overlapping date ranges when creating the temporal network. This would also have some other advantages. Soldiers were not

always able to write every day. For example, many of our authors who landed at Gallipoli on 25 April 1915 waited a few days to write up their account of the landing. Furthermore, overlapping date ranges would allow for links between people at the event and those who heard about it a few days later.

This could also be taken a step further when finding diaries of interest, by calculating the similarity between all pairs of entries, regardless of date. When looking for a diary of interest we want one which covers the experiences written in as many other diaries as possible. There would have been time differences in terms of when the soldiers embarked, were injured, took leave, and demobilized. Therefore, considering the similarity between all pairs of entries would take this into account.

Finally, the sentiment scores of the entries could be used to help find dates of interest. This may be in the form of finding days with high connectivity and widely varying sentiments, or days with two highly connected components with different average sentiments. However, before this can be done, a more accurate sentiment analysis technique would need to be used. For this paper, VADER was used to determine sentiment as it is easy to implement. This is a dictionary-based method that was developed for use with modern Twitter data. As such its dictionary may not be suitable for historic war diaries. Furthermore, it does not consider the context of a word, which can affect its sentiment. In future, it would be beneficial to find and apply a more accurate sentiment analysis technique to determine if this might be useful in finding dates of interest.

7. Conclusion

As the number of easily accessible transcribed historical documents increases, so does the need for methods to computationally analyze these documents. However, it is important to not eliminate traditional close reading analysis. Instead, we must find computational methods, which have the advantages of quickly analyzing many documents and viewing overall patterns, which can help locate representative documents for close reading analysis. This idea formed the basis of this paper, which investigated whether temporal network analysis can be used to analyze historical diaries. More specifically, the proposed methodology aimed to find dates and diaries of interest by considering networks of similar diary entries, as defined by cosine similarity. By applying the proposed methodology to a collection of 504 Australian World War I diaries held by the State Library of New South Wales, we were able to show that this method has potential.

However, there are several elements of the workflow that need further research. Our method used BERT to convert entries into a vector representation, and cosine similarity to determine the similarity of entries. This method was chosen as both BERT and cosine similarity are easy to implement and widely used. However,

there are other ways to calculate the similarity of documents, and these should be investigated to determine if the method used affects the results. Furthermore, a more comprehensive study into determining a similarity threshold should be undertaken. Due to the inaccuracy of automatically extracted dates in our diaries, it would be valuable to consider other ways of creating our network, including investigating similar entries within date ranges. Finally, sentiment analysis could be more fully included in the methodology.

Bibliography

Primary Sources

"Records of the European War Collecting Project, established by the Trustees of the Mitchell Library in 1918." State Library of New South Wales, https://collection.sl.nsw.gov.au/record/9arGy32n.

"The AIF Project." UNSW Canberra, https://www.aif.adfa.edu.au/index.html.

Secondary Sources

"The AIF Project." *Sabertache* 34, no. 2 (1993): 34–6.

Adams, Jennifer, and Kevin Gunn. "Digital Humanities: Where to Start." *College & Research Libraries News* 73, no. 9 (2012): 536–69. https://doi.org/10.5860/crln.73.9.8831.

Ahnert, Ruth, and Sebastian Ahnert. *Tudor Networks of Power.* Oxford: Oxford University Press, 2024. ProQuest Ebook Central.

Al-Anzi, Fawaz, and Dia AbuZeina. "Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing." *Journal of King Saud University – Computer and Information Sciences* 29, no. 2 (2017): 189, 191. https://doi.org/10.1016/j.jksuci.2016.04.001.

Atçil, Abdurrahman, and Gürzat Kami. "Studying Professional Careers as Hierarchical Networks: A Case Study on the Careers of Chief Judges in the Ottoman Empire (1516–1622)." *Journal of Historical Network Research* 7, no. 1 (2022): 1–32. https://doi.org/10.25517/jhnr.v7i1.113.

Australian War Memorial. "Enlistment statistics, First World War." Last modified March 18, 2022. https://www.awm.gov.au/articles/encyclopedia/enlistment/ww1.

Bean, Charles. "The Voyage and the Emden." In *The Story of ANZAC: From the outbreak of war to the end of the first phase of the Gallipoli Campaign, May 4, 1915 (11th edition).* Sydney: Angus & Robertson, 1941, 101–14. https://www.awm.gov.au/collection/C1416974.

Buzzetti, Dino. "The Origins of Humanities Computing and the Digital Humanities Turn." *Humanist Studies & the Digital Age* 6, no. 1 (2019): 32–58, https://doi.org/10.5399/uo/hsda.6.1.3.

de Arruda, Henrique, *et al.* "Representation of texts as complex networks: a mesoscopic approach." *Journal of Complex Networks* 6, no. 1 (2018): 125–44, https://doi.org/10.1093/comnet/cnx023.

de Arruda, Henrique, *et al.* "Paragraph-based representation of texts: A complex networks approach." *Information Processing & Management* 56, no. 3 (2019): 479–94. https://doi.org/10.1016/j.ipm.2018.12.008.

Dennis, Peter, *et al.* "Rosenthal, Major-General Sir Charles." In *The Oxford Companion to Australian Military History, 2nd Edition.* Oxford University Press, 2009. https://www.oxfordreference.com/display/10.1093/acref/9780195517842.001.0001/acref-9780195517842-e-997.

Devlin, Jacob, *et al.* "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,* Minneapolis, 2019.

Edmonds, Elise. "Of sentimental value: collecting personal diaries from the First World War." *Archives and Manuscripts* 48, no. 2 (2020): 186–99, https://doi.org/10.1080/01576895.2020.1755701.

El Gohary, Amira, *et al.* "A Computational Approach for Analyzing and Detecting Emotions in Arabic Text." *International Journal of Engineering Research and Applications (IJERA)* 3, no. 3 (2013): 100–7.

Erickson, Bonnie. "Social Networks and History: A Review Essay." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 30, no. 3 (1997): 149–157, https://doi.org/10.1080/01615449709601182.

Fernando, Samuel, and Mark Stevenson. "A Semantic Similarity Approach to Paraphrase Detection." In *Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics,* 2009.

Fields, Sam, *et al.* "Using named entity recognition and network analysis to distinguish personal networks from the social milieu in nineteenth-century Ottoman-Iraqi personal diaries." *Digital Scholarship in the Humanities* 38, no. 1 (2023): 66–86. https://doi.org/10.1093/llc/fqac047.

Gammage, Bill. *The Broken Years: Australian Soldiers in the Great War.* Ringwood: Penguin Books, 1990.

Gilbert, Alan, John Robertson, and Roslyn Russell. "Computing Military History: A Research Report on the First AIF Project." *War & Society* 7, no. 1 (1989): 106–13, https://doi.org/10.1179/war.1989.7.1.106.

Greteman, Blaine. *Networking Print in Shakespeare's England: Influence, Agency, and Revolutionary Change.* Redwood City: Stanford University Press, 2021. ProQuest Ebook Central.

Han, Jiawei, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques.* San Diego: Elsevier Science & Technology, 2011. ProQuest Ebook Central.

Hill, A. "Rosenthal, Sir Charles (1875–1954)." *Australian Dictionary of Biography: Volume 11, 1891–1939.* Australian National University: 2006. https://adb.anu.edu.au/biography/rosenthal-sir-charles-8268.

Hutto, C., and Eric Gilbert. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." In *Proceedings of the International AAAI Conference on Web and Social Media* 8, no. 1, 2014: 216–25, https://doi.org/10.1609/icwsm.v8i1.14550.

Janicke, Stefan, *et al.* "On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges." *Eurographics Conference on Visualization,* 2015, 83–103.

Jockers, Matthew. *Macroanalysis: Digital Methods and Literary History.* Champaign: University of Illinois Press, 2013. ProQuest Ebook Central.

Mardatillah, Ulfa, et al. "Citation Analysis on Scientific Articles Using Cosine Similarity." In Proceedings of the *2021 7th International Conference on Wireless and Telematics (ICWT),* Bandung, 2021. https://doi.org/10.1109/ICWT52862.2021.9678402.

Mickus, Timothee, *et al.* "What do you mean, BERT? Assessing BERT as a Distributional Semantics Model." *Society for Computation in Linguistics* 3, no. 1 (2020): 350–61, https://doi.org/10.7275/t778-ja71.

Mihalcea, Rada, Courtney Corley, and Carlo Strapparava. "Corpus-based and Knowledge-based Measures of Text Semantic Similarity." in *Proceedings of the 21st national conference on Artificial intelligen*ce, Boston, 2006, 775–80.

Newman, Mark. *Networks, 2nd Edition.* Oxford: Oxford Academic, 2018. https://doi.org/10.1093/oso/9780198805090.001.0001.

Olson, Wes. *The Last Cruise of a German Raider: The Destruction of SMS Emden.* Pen and Sword, 2018.

Robson, Lloyd. "The origin and character of the first A.I.F., 1914–1918: Some statistical evidence." *Historical Studies* 15, no. 61 (1973): 737–49. https://doi.org/10.1080/10314617308595502.

Stein, Evina, and Gustavo Fernández Riva. "Introduction: Fitting Manuscript Studies into the Historical Network Research." *Journal of Historical Network Research* 9, no. 1 (2023): iii–xvi. https://doi.org/10.25517/jhnr.v9i1.

Zimmer, Ben. "Stealth Words." review of *The Secret Life of Pronouns: What Our Words Say About Us,* by James Pennebaker. *The New York Times Book Review,* August 28, 2011. Gale Academic OneFile.