Author Declaration for Group Assignments

Title of Assignment: Group 4 Final Essay
Module Number: CS7IS4
Word Count: 4,512
Date: April 14th 2024

Contributions

1 Contributions

Student ID	Full name	Role	Nature of Contribution
19306182	Mingwei Shi	Chair	1. Write interpretation report 2. Communicate with Prof.Carl Vogel to confirm topics 3. Implement email graph modelling 4. Write the abstract, introduction, introductory paragraph of related works and methodology, the introductory paragraph of dataset construction (3.1), and dataset comparison (3.3) in the middle stage report. 5. Chair and host weekly meetings and arrange tasks for each member 6. Conduct experiment on email dataset 7. Write an experiment related to the email dataset 8. Update abstract 9. Read peer reviews from other groups and summarise it into a group spreadsheet. 10. Participate in writing the conclusion section
23341031	Juliette van Marken	Recorder	1. Write related work section of report 2.1 and 2.2 2. Create 5 summaries of research articles that were used for related research 3. Helped conceptualize methodology approach 4. Created minutes for every meeting 5. Write future work section
23336943	Akash Garg	Accountant	1. I worked on the Email Data set and its visualization 2. Write and research about “related work”, "discussion” and clustering in the midterm report. (Sections 2.1 / 2 and 5) 3. Make Topic related, 3 Research Article summary 4. Read peer reviews from other groups and summarise them into a group spreadsheet. 5. Handle Accountant Duty
23335292	Georgios Kanellopoulos	Ambassador	1. Reviewed interpretation report. 2. Implemented the research article citations graph. 3. Contributed to the methodology paragraph of the midterm report (section 3.2). 4. Summarized 2 research articles that were used for the literature review of the midterm report. 5. Attended 4 meetings of other groups and shared ideas taken from these meetings in ambassador reports. 6. Summarized peer reviews from other groups. 7. Contributed to the email and citation dataset experiment. 8. Contributed to the abstract, methodology and implementation sections of the final report. 9. Reviewed the final report.
23337042	Xinyi Li	Monitor	1. Improve interpretation report 2. write 3.1(3.1.1&3.1.2) for the midterm report 3. make 3 related summary for the topic 4. read and summary other groups’ peer reviews

			5. write “discussion”(5) part of the final report 6. handle the monitor duty
23345983	Indrajeet Kumar	Verifier	1. Contributed to the Graph construction methods sub-paragraph of the midterm report (Section 3.1.1 and Section 3.1.2) 2. Created 1 summaries of research articles that were used for related research 3. Handled the verifier duty 4. Added research hypothesis including positive and false hypothesis in the introduction section 5. Read and Summarise other groups’ peer reviews 6. Engage in the conclusion writing

1 Declaration

We have read and we understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at: http://www.tcd.ie/calendar.

We have also completed the Online Tutorial on avoiding plagiarism, ‘Ready, Steady, Write’, located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write.

We declare that this assignment, together with any supporting artifact, is offered for assessment as our original and unaided work, except in so far as any advice and/or assistance from any other named person in preparing it and any reference material used are duly and appropriately acknowledged. We declare that the percentage contribution by each member, as stated above, has been agreed by all members of the group and reflects the actual contribution of the group members.

2 Signatures

Signed and dated:

1. Mingwei Shi Mingwei Shi 2024, April 14th
2. Juliette van Marken 2024, April 14th
3. Akash Garg Akash Garg 2024, April 14th
4. Kanellopoulos Georgios 2024, April 14th
5. Xinyi Li Xinyi Li 2024, April 14th
6. Indrajeet kumar Indrajeet 2024, April 14th

Exploration of Textual Similarities in Network Communication

Mingwei Shi, Juliette Van Marken, Akash Garg ,
Georgios Kanellopoulos, Xinyi Li, Indrajeet Kumar
School of Computer Science and Statistics, Trinity College Dublin
{mshi, vanmarkj, akgarg, kanellog, lix13, kumari}@tcd.ie

Abstract

Individuals chatting with each other often share similar phrases and mantra. To explore this phenomenon on social networks, we explored textual similarities and word patterns in two communication graphs: a graph of citations between scientific publications and a graph of emails exchanged between Enron company employees. We calculated the degree of overlap between the network of communications and the network of similarity of communications for every graph, and we examined word patterns and sentiments in the intersection set of these networks that show to what extent communication and textual similarity overlap. Although the results do not show a clear relationship between text similarity and communication, there are indications that people who communicate often tend to use same phrases and share the same sentiment. These findings are valid for the email exchange dataset, but could not be verified for the scientific publications dataset. We explored further the findings for the email dataset and found that sentiments are not correlated to the amount of overlap, while text semantics are positively correlated.

Keywords— network of communications, network of similarity of communications, social network analysis, sentiment analysis, semantic similarity, text mining

1 Introduction

Our study starts with the idea "Birds of a feather flock together", suggesting that people who often talk or interact with each other are likely to talk in a similar manner. This idea leads us to examine how these patterns manifest within networks, specifically looking at the intersection between individuals' social networks—whom they communicate with—and similarity networks-what they communicate.

Our research seeks to answer a key question: **" To what extent are there similar text and word patterns in network communication?"** To explore this, we use two different datasets: one from the Enron email corpus and another representing scientific paper citations. These datasets help us trace the threads of conversation and idea exchange to see how closely connected individuals align in their use of language and choice of topics.

Two main ideas are tested:

- The first idea is that there is a significant overlap between the networks based on whom individuals communicate with and the language the individuals use. This hypothesis suggests that regular interaction within these networks leads to a convergence in the language used, mirroring the principle that similar individuals tend to group together. for example, in a company like Enron or in academic scholars represented by the citation network, we expect to find clusters of individuals whose communications are not just frequent but also topically and linguistically coherent. This would imply that the social ties people form significantly influence the ideas they share and the manner in which they express these ideas.
- The second idea is the opposite, we consider the possibility that no significant overlap exists between social networks and similarity networks. This would challenge the old saying that "birds of a feather flock together". It means that even in places where you'd expect people to be on the same page, like among Enron's staff or in academic circles, there's still a lot of variety in what people are interested in and how they like to talk about it. This outcome would indicate that other factors, perhaps external to the network itself, play a more pivotal role in shaping communication patterns.

We have organized our paper to take a thorough look at these ideas. Section 2 reviews the previous literature in relation to our research. Section 3 points out the methodology concerning the construction of the dataset, analysis methods applied in the dataset and comparison between the two datasets. Section 4 presented the experiment results. The last three sections exhibit discussion, conclusion, and future work.

2 Related works

The Related Works section grouped our literature review into two categories.The first category,"Social communication within the graph", aims to construct research territory to readers, provide the purpose of our research, and identify the research gap from existing work,such as Healey et al. (2007). The second category,"Text and word pattern similarity research", focuses on the evaluation metrics in previous works and the rationale for choosing these measurements as our evaluation metrics.

2.1 Social communication within a community

As mentioned in the introduction, it is hypothesised that people who communicate often, tend to be more similar. To this end, Brinberg & Ram (2021) did research into the linguistic similarity of a romantic couple. They found that over time, the linguistic similarity, which was measured of text messages, increased. These findings give strength to the idea that the individuals you communicate with have an effect on the linguistics you use. Healey et al. (2007) have done research into the idea that people in the same community tend to be more similar by creating a similarity set and a contact set to take a look at the degree of overlap between the two sets. In their research, they used chat logs of an online social community with over 1500 regular users. They found that a large amount of people interacted with people that were dissimilar to them. This finding revealed that, at least for this dataset, language convergence suggests that direct communication is not the main factor influencing convergence. It is important to note that this data contained direct communication without the presence of an influential individual.

Ikeda et al. (2013) aimed to use a users social media presence to estimate their demographic. A model was trained on $100,000$ twitter users in order to attempt to determine the demographic. From these users, the tweet content and social network, meaning followers and following, were extracted. Based on the extracted text and community, they were able to make an effective estimate for occupation, age, area and hobby. The ability to use tweets, which are indirect communication, and communities to estimate the demographic of a user, gives confidence to the idea that there could possibly be a correlation between language use and community.

Where previous research has focused on informal communication with high velocity of communication, our research aims to look into whether similar patterns can be observed in formal communications with a lower velocity of communication. In the context of this research, 'velocity of communication' refers to the frequency of communication and the time passed between communications.

our research contextualizes the notion of linguistic similarity within social communication networks. Building upon previous studies such as Healey et al. (2007), which explored linguistic convergence within online communities, our research aims to extend these findings to formal communication settings with a lower velocity of communication. We recognize the importance of clarifying the concept of "velocity of communication" and its relevance to our investigation. By focusing on formal communication contexts, we aim to contribute insights into linguistic convergence beyond informal interactions, thereby enriching the understanding of social communication dynamics.

2.2 Text and word pattern similarity research

When attempting to calculate the similarity between two text, it is important to determine what kind of similarity you aim to measure. Words can be similar in two different ways, lexical and semantically(Vijaymeena & Kavitha, 2016). Lexical refers to the sequence of characters used, while semantically refers to the theme of the word. When we want to find two documents that cover similar topics, for example when looking for references, we want to find documents that are semantically similar. This is a common practice used in the field of information retrieval. On the other hand, finding lexical similarity is more common in the field of text analytics. Lexical information gives us an insight into the language that was used.

A common measure when trying to find lexical similarity is n-grams. In previous research, n-grams were used in an attempt to find the similarity between dialects in an online community Healey et al. (2007). N-grams, which split up a text in sub-strings of n characters, have the benefit of being able to look at sub-words, which can allow for the comparison of words that are differently inflected forms of the same root. Healey et al. also introduced the idea that even uni-grams, which look at individual letters, could give valuable information about the language used by looking at the frequency of certain letters, such as 'Q', being used.

A measure that is commonly used to measure the similarity for clustering documents is TF-IDF Bafna et al. (2016). TF-IDF, which stands for Term Frequency - Inverse Document Frequency, takes the frequency of a word into account, compared to the length of a document, to determine the importance of that word. TF-IDF is often used in combination with cosine distance matrix in order to measure how similar two documents are. A limitation of TF-IDF is that, unlike n-grams, two words that are different forms of the same root, will be seen as different words. However, this limitation can be avoided by doing pre-processing, as seen by Patil & Atique (2013). Introducing stemming, which reduces a word back to its root form, will avoid two different inflected forms of the same root being seen as different words.

In addition to using methods such as n-grams and TF-IDF to evaluate text similarity, the Jaccard similarity measure can also effectively analyze the correlation between texts. Salvatore et al. (2020) Research demonstration in the field of bioinformatics This paper demonstrates the potential of using the Jaccard index to evaluate gene expression network similarity. It shows that although the principle of Jaccard similarity is simple, it has shown its unique ability in revealing network structure and exploring similarities. However, it should be noted that Jaccard similarity sensitivity to data set size may have an impact on analysis results and needs to be considered when applying this.

We further clarify the distinction between lexical and semantic similarity in text analysis. Building on previous research by Healey et al. (2007) and Vijaymeena & Kavitha (2016), we acknowledge the importance of understanding different measures of text similarity, including n-grams and TF-IDF. We recognize the limitations of TF-IDF in capturing variations of word forms and emphasize the significance of preprocessing techniques such as stemming, as highlighted by Patil & Atique (2013). Additionally, we acknowledge the potential of alternative measures such as the Jaccard similarity index, as demonstrated by Salvatore et al. (2020), particularly in revealing network structure and exploring similarities. We also note the sensitivity of Jaccard similarity to dataset size, which warrants consideration in our analysis.

3 Methodology

The followed methodology can be split into three major component: graph construction from each dataset, a unified graph analysis methods and comparisons between both datasets. The first step is to transform the datasets into a graph structure with nodes and edges, and allocate the text to the nodes. Subsequently, the networks of communication and network of similarity of communications is defined for every node in the graph. The research goal is to calculate the amount of overlap between these networks and compare the results between different datasets (cf.figure.3.1).

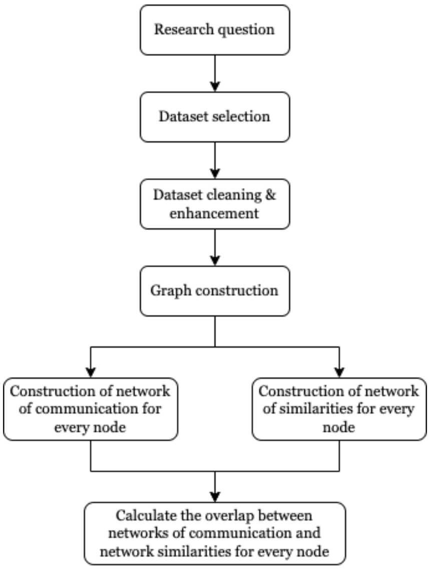
Research question
Dataset selection
Dataset cleaning & enhancement
Graph construction
Construction of network of communication for every node
Construction of network of similarities for every node
Calculate the overlap between networks of communication and network similarities for every node

Figure 3.1: The methodology diagram

3.1 Graph construction methods

This research uses two different datasets: the Enron email dataset, which contains records of email exchanges between employees; and the scientific citation dataset, which contains a range of information about the paper including the abstract and citation relationships. The following sections describe the methodology followed to construct graphs from these two datasets.

3.1.1 Enron corporation email dataset

In the preprocessing stage of the Enron email dataset, we used several functions from various websites to clean and prepare the data for analysis. We used the 'get_text', 'get_row', and 'get_address' functions from a Kaggle article[1] in our scripts to extract the important information from the emails, such as the email body text, the sender and receiver's email addresses, and the date.

To ensure the consistency and clarity of the data, we used the "standard_format" function from the same article[2] to clean and standardize the data. This step was crucial as it removed a total of 111,433 data entries that were missing essential information and were therefor determined to be not suited for the research.

For emails that were sent to multiple recipients, we used the 'split_email_addresses' function from this website website[3] to separate the email addresses. This ensured the accuracy of the interactions in the communication network.

In the communication network created from the email dataset, the senders and receivers of emails represent the nodes, while the communication relationships are represented by the edges. To focus on direct communication and improve the analysis of interaction patterns, we excluded group emails and forwarded emails. This way, we can better understand how individuals communicate with each other and create better clusters in our future work. Lastly, we transfer the text between senders and receivers to the senders node in order to fit our research topics.

3.1.2 Scientific paper citations dataset

The scientific paper citations dataset[4] comprises of two JSON files. The citation_relations.json file contains information about the citation relations between different papers, and the papers.SSN.jsonl file contains paper information, such as title, authors, abstract and body.

[1] https://www.kaggle.com/code/oalvay/enron-emails-complete-preprocessing
[2] https://www.kaggle.com/code/oalvay/enron-emails-complete-preprocessing
[3] https://www.kaggle.com/code/gpreda/parse-and-process-enron-emails-dataset
[4] https://github.com/ChenxinAn-fdu/CGSum

During the pre-processing phase, we checked for duplicates to make sure that there is no duplicate information. Then, we transformed the data into a format that can be used to create a network with Python's `NetworkX` library. The required format is a list of dictionaries with source and target paper identifiers for every edge.

A node in this context represents a single paper, and an edge represents a citation link between two papers. By connecting these nodes with edges based on the citation relationships, we created a directed citation graph, allowing us to analyse the structure and properties of the citation network. Every node also contains the abstract of the relevant paper. We chose to work with the abstract only to make our calculations resource-efficient.

The original dataset contains more than 161K nodes and 660K edges. As the processing of such a large dataset would require more processing power than available to us, we decided to sample the graph and work with a smaller amount of data. This was achieved with the ”Little Ball of Fur” graph sampling library Rozemberczki et al. (2020). To maintain the properties of the graph, we preferred a degree-based node sampling method instead of random sampling Adamic et al. (2001).

3.2 Graph analysis methods

Graph analysis methods incorporate three components: the construction of networks of communications, the construction of networks of similarities,and the calculation of overlap. This process has been visualised in figure 3.2 and the steps will be explained in more detail in the following subsections.

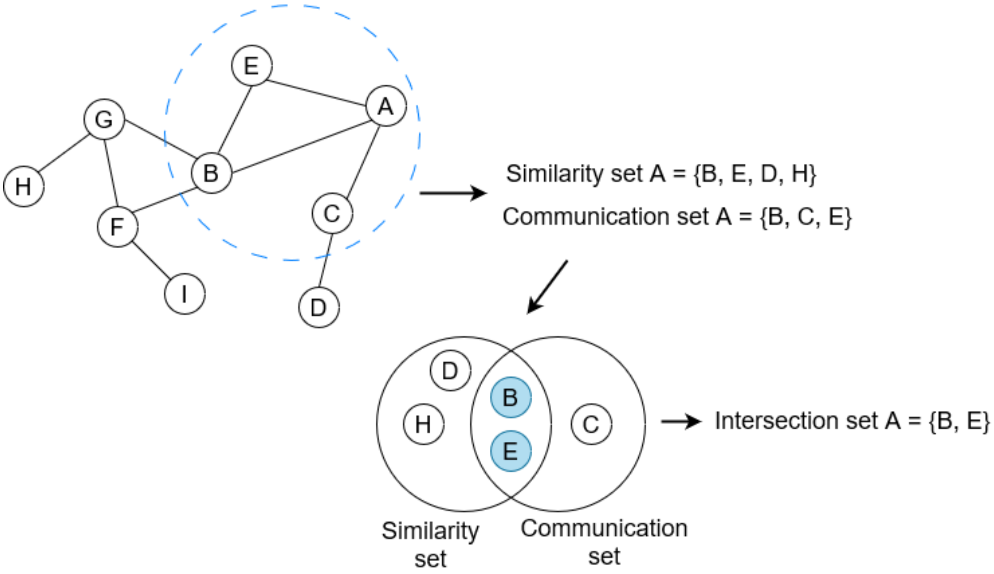
E
A
G
B
H
C
F
I
D
Similarity set A = {B, E, D, H}
Communication set A = {B, C, E}
D
B
H
E
C
Intersection set A = {B, E}
Similarity set
Communication set

Figure 3.2: Graph analysis process

3.2.1 Construction of networks of communications

The network of communications is calculated for every node in the graph and consists of its adjacent nodes. In other words, the network of communications is the papers that an article has referenced directly or the colleagues that an employee has sent an email to. If we refer back to figure 3.2, which shows an example for node 'A', the resulting network of communication would contain 'B', 'C' and 'E', as these are all the nodes that are directly connected to 'A'.

3.2.2 Construction of networks of similarities

The graph of similarity of communications connects nodes whose messages display text similarity. The following steps are followed to build graphs of similarity of communication. Firstly, text information is extracted from every node in the graphs. Secondly, n-gram vectorisation (taking into consideration unigrams, bigrams and trigrams and their combinations) is applied to represent every node's text as a vector in a high-dimensional space. A n-grams approach is chosen because it allows for understanding text semantics and it is not topic-discriminating. Finally, cosine similarity is calculated between all node pairs. Edges are drawn between nodes whose pairwise similarity exceeds a threshold, and the weight of the edge is the value of the similarity metric.

The selection of an appropriate threshold is necessary to prevent the graph of similarities from becoming too dense and noisy. For this reason we calculated the distribution of all pair-wise text similarities and draw an edge when the similarity falls in the top 10 percent of the distribution.

3.2.3 Calculation of overlap

Having calculated the network of communications and the network of similarity of communications for every node, the degree of overlap is determined by the intersection of nodes in the two networks. For the citations graph, a total overlap means that a scientific paper cites only paper that have significant text similarity with it. Equally, for the email graph it means that an employee has exchanged emails only with people who produce text significantly similar to theirs. This calculation has been visualised in figure 3.2 with a Venn diagram, as can be seen in the figure, the overlap can be found by taking the intersection of the Venn diagram.

3.3 Comparison between two datasets

Given the two datasets, we can state the differences between the datasets in terms of communication, including two aspects (interaction quantity and interaction update frequency). In one aspect, there might be multiple interactions between senders and receivers of emails, whereas there is only one interaction possible between two scientific papers. From another aspect, multiple communications within the Enron Email dataset can happen in a short time period and relatively fast email replies are to be expected due to the nature of email communications within a corporation. On the contrary, there can be a long time period between the publication of the two original paper and the publication of a paper that cites it.

4 Experiment Result and evaluation

Table.4.1 presents information on node count, edge count, average node degree, and median node degree for the email and citation datasets. Notably, the email dataset is relatively smaller than the citation dataset based on node count and edge counts(28:1291; 149:1593), while the email dataset is more dense than the citation dataset based on average node degree and median node degree, as these two metrics reflect the density of each node flow.

Dataset	Nodes count	Edges count	Average node degree	Median node degree
Emails	28	149	11.5	12
Citations	1291	1593	2.5	2

Table 4.1: Network of communications statistics for both datasets

Based on this fact, this work explores a different range of n-grams to test associated attributes in the scientific article dataset and Enron corporation email dataset.

4.1 Scientific articles dataset

n-gram range	Node count	Edge count	Average node degree	Median node degree
1	1087	83333	153	90
2	1277	83334	130	76
3	1273	83330	130	81
[1,3]	1043	83334	160	76
[2,3]	1277	83334	131	76

Table 4.2: Network of similarities statistics for different n-gram ranges (scientific citations dataset

The experiments on the scientific citations graph dataset aim at quantifying the amount of overlap between the network of communications and the network of similarity of communications for every node in the graph. After

n-gram range	Jaccard index	p-value
1	0.018	0.64
2	0.008	0.57
3	0.002	0.77
[1,3]	0.010	0.69
[2,3]	0.006	0.62

Table 4.3: Permutation test results for different n-gram ranges (scientific citations dataset)

Jaccard similarity is calculated for these two networks for every node, a permutations test is carried out to determine the statistical significance of the findings. In the permutation test, the network of communication of a specific node is selected and the Jaccard similarity is calculated with respect to the network of similarity of communications of 100 randomly selected nodes, and then the p-value is calculated. This procedure is repeated 100 times. The null hypothesis of the test is that there is no overlap between the two networks ($p > 0.05$).

Table 4.2 reports selected statistical properties of the networks of similarity of communications for different n-gram ranges. As the n-gram range is increased from unigrams to trigrams, more nodes are included in the network of similarities of communications. In this dataset, the choice n-gram range does not seem to change the number of edges that are over the text similarity threshold. Results show that there is no significant amount of overlap between the network of communications and the network of similarity of communications for a specific node. Our experiments results are presented on table 4.3 and show that Jaccard similarity between the network of communications and network of similarity of communications is between 0.002 and 0.01. The p-value for the permutation tests is between 0.57 and 0.69, thus proving that most random pairs of networks of communications and networks of similarity of communication have a higher Jaccard similarity. Therefore, the null hypothesis of the test is verified.

It is worth noting the great difference in the average and median node degree between the network of communications (table 4.1 and the networks of similarity of communications 4.2. The network of communications is a sparsely connected graph with an average node degree of 2.5, while the network of similarities has much more edges and the average node degree is over 130 for all n-gram configurations. This asymmetry in node degree inevitably leads to a low similarity between the two networks.

4.2 Email dataset

We repeated the same experiments for the email dataset. The statistical properties of the networks of similarities of communication for this dataset are presented in table 4.4. It is remarkable that unigrams lead to a network of similarities with more edges and a higher node degree that higher order n-grams. This observation infers that the text similarity is more uniformly distributed when unigrams are used.

The results of the permutation test appear on table 4.5. The highest p-value is observed when unigrams, bigrams and trigrams are taken into consideration. Even in this case the p-value is not high enough to reject the null hypothesis of the permutation test.

n-gram range	Node count	Edge count	Average node degree	Median node degree
1	26	98	7.5	8
2	19	39	4.1	3
3	20	39	3.9	4
[1,3]	21	39	3.7	3
[2,3]	20	39	3.9	3

Table 4.4: Network of similarities statistics for different n-gram ranges (email dataset)

To explain the results further, we conducted a sentiment and semantic similarity analysis to explore the mechanism behind the overlap in networks of communications and networks of similarity of communications for the email dataset. To perform sentiment analysis, stop words were removed from the text of all nodes and lemmatisation was performed.

n-gram range	Jaccard index	p-value
1	0.42	0.16
2	0.33	0.23
3	0.43	0.14
[1,3]	0.50	**0.13**
[2,3]	0.40	0.19

Table 4.5: Permutation test results for different n-gram ranges (email dataset)

The Enron email dataset contains emails that were exchanges prior to the corporation's bankruptcy in 2003. Therefore, it is reasonable that staff had negative feelings during that time. As shown in figure 4.1, negative emotions are indeed more prevalent for all email senders. However, the percentage of negative emotions is only marginally larger than positive counterparts, as shown in figure 4.2.

n-gram range	Sentimental correlation	Semantic correlation
1	0.10	0.95
2	0.27	0.95
3	-0.26	0.96
[1,3]	-0.04	0.96
[2,3]	0.05	0.95

Table 4.6: Pearson correlation values between the Jaccard similarity, the negative sentiments score and the semantic similarity scores for different n-gram ranges (email dataset)

Table 4.6 shows the results of the correlation between the negative sentiments score and the Jaccard similarity for different n-gram ranges. The negative sentiments score is calculated as the percentage of negative words in a sender's 100 most frequent used used. This score is calculated for each intersection set and it is correlated with the Jaccard similarities. The results show that there is no significant correlation between the use of words with a negative sentiment and the amount of overlap between networks of communication and networks of similarity of communication. This finding is expected, given that positive and negative are almost equally used.

Then, we used a semantic analysis toolkit from "Spacy"[5] that incorporates synonym detection, synonym relationship detection and contextual phrase identification. This powerful tool could find semantic similarity between lemma vectors in each sender node. Figure 4.3 visualises the semantic similarity between all pairs of nodes. We can notice that all nodes have over 0.9 semantic similarity with other nodes, which explains the relatively high overlap between similarity and communication networks in the email dataset.

Table 4.6 shows the correlation between the average semantic similarity and the Jaccard similarity in the intersection sets. There is a clear positive correlation between these two metrics, indicating that a higher degree of overlap between a network of communications and a network of similarity of communications is correlated with high semantic similarity. In other words, the use of common n-grams between nodes is correlated with semantic similarity.

5 Discussion

In this study, we examined the textual similarities and word patterns in network communication, focusing on two distinct datasets: the Enron email corpus and a dataset representing scientific paper citations. Our analysis aimed to shed light on the convergence or divergence of communication and similarity networks, providing insights into how individuals within networks align in their language use and choice of topics.

Our investigation revealed some interesting linguistic patterns within the Enron email dataset and the scientific paper citations dataset. Notably, we found varying levels of overlap between communication and similarity networks in the two datasets. While the Enron email dataset exhibited relatively high overlap, suggesting some amount of convergence between communication and topical similarity, the scientific paper citations dataset showed fewer

[5] `https://spacy.io/`

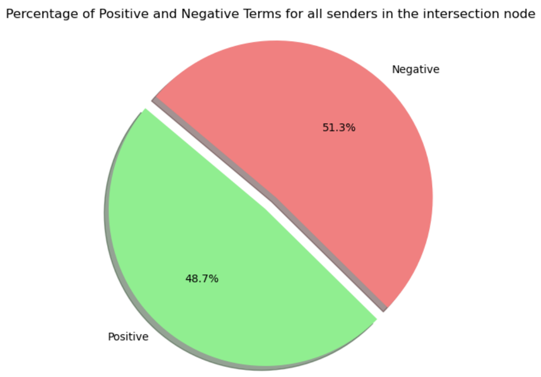

Figure 4.1: Sentimental polarity on all senders in the intersection node

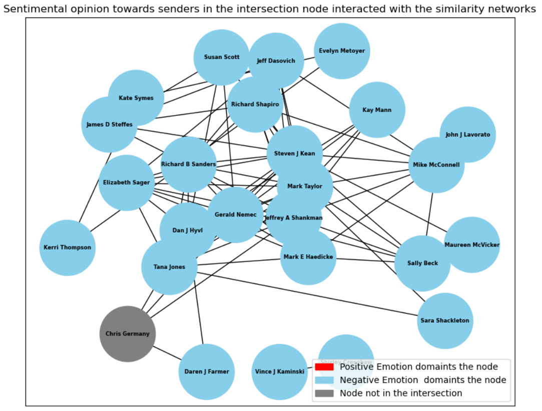

Figure 4.2: The dominant emotion in each node in the similarity networks

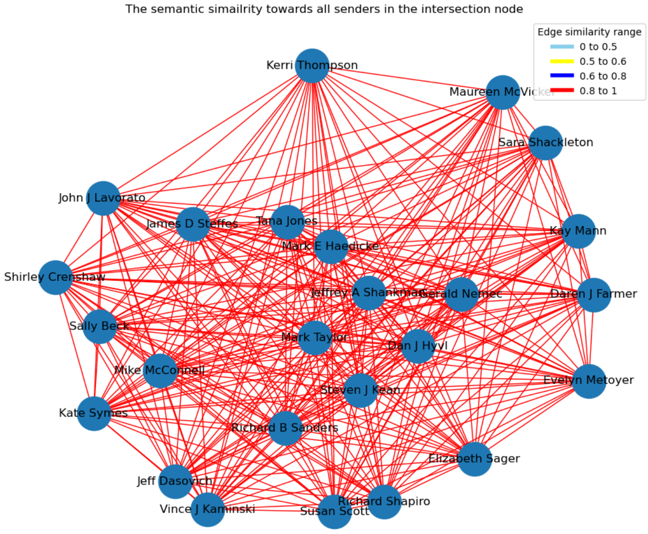

Figure 4.3: Semantic similarity between each senders in the intersection node

overlaps, indicating weaker linguistic ties among peer-reviewed scientific articles.

Our findings align only moderately with previous research on social communication within networks and text similarity analysis. Contrary to the expectation that the degree of overlap will be high, based on prior studies, we found that there is no statistically significant amount of overlap for the citations dataset. Our findings could indicate that indirect communication that happens after a long time is not related to using similar texts. However, further research is needed to verify this argument; our methodology used sampling to reduce the size of the citations graph, thus skewing its statistical properties significantly (mainly due to the removal of a large number of edges).

Results are closer to expected for the emails dataset. Experiments establish a relatively high overlap between networks of communications and networks of similarities and correlate this overlap with semantic text similarity. Sentiment does not seem to influence the amount of overlap between these networks. This finding can be supported by the observed phenomenon of rapid formation of group sublanguage in online communities Healey et al. (2007).

6 Conclusion

In conclusion, this work addresses the research question,“To what extent are there similar text and word patterns in network communication?” by validating two diverse datasets, the Enron Corporation email and scientific article datasets. Due to differences between the two datasets, including capacity, connectivity, activity, and the level of richness of text information, we considered different approaches to constructing the two networks. We adopt the same analysis for two graphs to control the condition by identifying the overlapping between communication and similarity networks.

From our research, we found quite a contrast. In the Enron emails, there was a lot of overlap. This suggests than in a corporate environment, people tend to reuse some phrases, possibly due to the established communication protocol. On the other hand, in the scientific articles, there was not much overlap. This shows that in the academic world, even when papers cite each other, they do not necessarily talk about the same things or use the same words. This could be because academic work covers a wide range of topics and ideas.

Overall, our study shows that the amount of similar topics and word patterns in network communication can vary greatly depending on the context. In a corporate environment like Enron, people tend to talk more similarly, while in academia, there's a lot more diversity in communication. This finding opens up more questions about how different settings influence the way we communicate.

7 Future Works

In this paper we discussed the results for two different data sets. Although both data sets have some characteristics in common, such as formal language being used, they are quite different in a few aspects. As previously described, two main differences between the two datasets are the quantity of interactions and the time passed between interactions. As we found different results for the two data sets, it would be interesting to find out how and to what extent each individual aspect contributed to this observed difference. Different data sets could be tested in order to analyse the impact of each individual factor.

In our research, we created a network of communication, which consisted of all the users a user had interacted with, but the quantity of messages between two different users was left out of scope. Future research could be done into whether there is a correlation between the frequency of communication and similarity. As previous research into the connection between communication and similarity focused on individuals that often interacted with each other, it could be interesting to see whether less frequent communication would align with lower similarity between individuals.

Where previously research often concerned itself on informal language, such as text taken from public forums, the two data sets used in this research contain mainly formal language. Future research could be done into comparing the two different kinds of language in order to see their impact on the observed similarity.

References

Adamic, L. A., Lukose, R. M., Puniyani, A. R., & Huberman, B. A. (2001). Search in power-law networks. *Physical Review E*, *64*(4).
URL `http://dx.doi.org/10.1103/PhysRevE.64.046135`

Bafna, P., Pramod, D., & Vaidya, A. (2016). Document clustering: Tf-idf approach. In *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*.

Brinberg, M., & Ram, N. (2021). Do new romantic couples use more similar language over time? evidence from intensive longitudinal text messages. *Journal of Communication*.

Healey, P. G., Vogel, C., & Eshghi, A. (2007). Group dialects in an online community. In *DECALOG 2007, The 10th Workshop on the Semantics and Pragmatics of Dialogue*, (pp. 141–147).

Ikeda, K., Hattori, G., Ono, C., Asoh, H., & Higashino, T. (2013). Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems*, *51*, 35–47.
URL `https://www.sciencedirect.com/science/article/pii/S0950705113002025`

Patil, L. H., & Atique, M. (2013). A novel approach for feature selection method tf-idf in document clustering. In *2013 3rd IEEE International Advance Computing Conference (IACC)*, (pp. 858–862).

Rozemberczki, B., Kiss, O., & Sarkar, R. (2020). Little ball of fur: A python library for graph sampling.

Salvatore, S., Rand, K. D., Grytten, I., Ferkingstad, E., Domanska, D., Holden, L., Gheorghe, M., Mathelier, A., Glad, I., & Sandve, G. K. (2020). Beware the jaccard: the choice of similarity measure is important and non-trivial in genomic colocalisation analysis. *Briefings in Bioinformatics*, *21*(5), 1523–1530.
URL https://doi.org/10.1093/bib/bbz083

Vijaymeena, M. K., & Kavitha, K. (2016). A survey on similarity measures in text mining. In *Machine Learning and Applications: An International Journal (MLAIJ)*.