Temporal Graph Analysis of Misinformation Spreaders in Social Media

Joan Plepi[*†] and **Flora Sakketou**[*†] and **Henri-Jacques Geiß**[‡] and **Lucie Flek** [†]
[†]Conversational AI and Social Analytics (CAISA) Lab
Department of Mathematics and Computer Science, University of Marburg
[‡]Department of Computer Science, Technical University of Darmstadt
{flora.sakketou, joan.plepi, lucie.flek}@uni-marburg.de,
henri-jacques.geiss@stud.tu-darmstadt.de
[*] These authors contributed equally to this work

Abstract

Proactively identifying misinformation spreaders is an important step towards mitigating the impact of fake news on our society. Although the news domain is subject to rapid changes over time, the temporal dynamics of the spreaders' language and network have not been explored yet. In this paper, we analyze the users' time-evolving semantic similarities and social interactions and show that such patterns can, on their own, indicate misinformation spreading. Building on these observations, we propose a dynamic graph-based framework that leverages the dynamic nature of the users' network for detecting fake news spreaders. We validate our design choice through qualitative analysis and demonstrate the contributions of our model's components through a series of exploratory and ablative experiments on two datasets.

1 Introduction

With the popularity of social media platforms constantly increasing, the dissemination of false online information becomes a major hurdle, having catastrophic effects on our society (McKay and Tenove, 2021). It is essential to address this issue early on; to efficiently and rapidly identify misinformation spreaders and spurious accounts which are likely to propagate posts from unreliable news sources. To this end, we introduce an early warning model that distinguishes authors who have repeatedly shared news from unreliable sources in the past, from those that share news from reliable sources. We use the terms 'misinformation spreaders' and 'real news spreaders' for each user class, respectively. In this paper, the term *misinformation* is used as an umbrella term that covers *misinformation, disinformation, partisan news and satirical content*. Figure 1 depicts examples of the posting activity for each user class.

Recently, significant attention has garnered towards graph representational learning methods (Wu et al., 2021) due to their advances in various NLP domains. Kim and Ko (2021) use a graph-based approach to model the semantic relationship between sentences in a document for fake news detection. Rath et al. (2021) apply graph neural networks to explore the social network of misinformation spreaders and show that interpersonal trust plays a significant role in differentiating them from real news spreaders. Such graph approaches are able to model user-to-user relationships and therefore provide a promising underexplored research direction for identifying misinformation spreaders.

The impact of time on fake news prediction has made the task even more challenging, as the content-based differences of news sources change due to the highly dynamic nature of the news topics (Horne et al., 2019). Most of the fake news detection methods that use static features need to be continuously updated with new annotated data to stay relevant (Kwon et al., 2017). We argue that this hypothesis can be generalized for detecting misinformation spreaders. Similarly to feature-based methods, existing graph modeling approaches are not specifically designed for learning the time-evolving similarities of the users' interactions. Addressing these limitations of existing research, we propose an approach accounting for the temporal dynamics of user-to-user relationships instead. We introduce a model that extracts features from users' content similarities and social interactions and models the temporal evolution of these connections in order to identify misinformation spreaders. In addition, our study aims to answer the following research questions:

RQ1: Do the users' semantic similarities and social interactions fluctuate over time?

RQ2: Are temporal relationships indicative of misinformation spreading behavior?

For the first exploration, we formulate the problem as a binary classification task, with a potential for a more fine-grained approach in the future. We

Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing, pages 89–104
October 16, 2022.

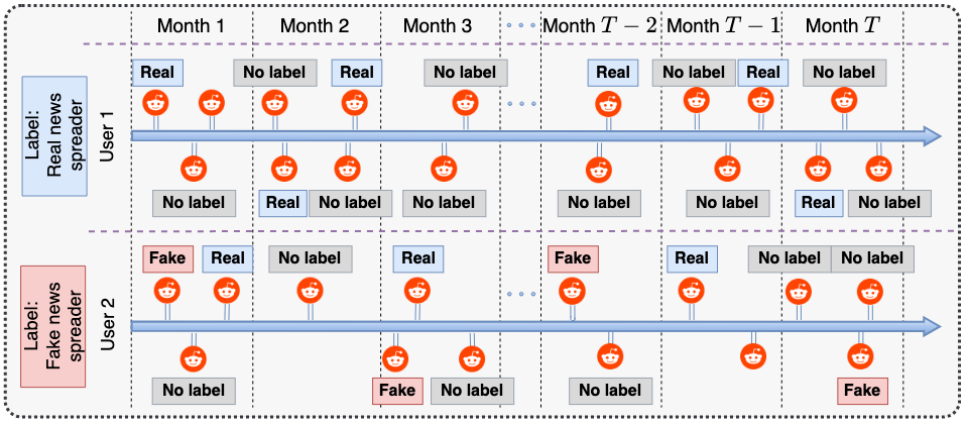


Figure 1: **Examples of the user classes.**

first build dynamic linguistic and social graphs, which are constructed based on the users' posting behaviour within consecutive time-windows. Subsequently, the generated temporal graph representations are treated as a sequence of features for the final classification. To the best of our knowledge, dynamic graph modelling has not been utilized for identifying misinformation spreaders in other works. We conduct a series of exploratory analyses in the user-to-user relationships. Through ablative experiments, we show the effectiveness of our model's components for profiling misinformation spreaders. Our contributions are as follows:

- We provide a comprehensive qualitative and quantitative analysis of the users' temporal semantic and social similarities and investigate the different types of dynamic graph connections.
- We develop a dynamic graph neural network framework for (a) predicting the users' future misinformation spreading behavior, (b) predicting the behavior of unseen users, and (c) predicting misinformation spreading behavior in a zero-shot scenario.
- We show that our proposed dynamic framework outperforms the baseline content-based models as well as the static graph model.
- We release our code to encourage future research.

2 Background and Related Work

While user profiling approaches have been investigated for various tasks, it wasn't until after the PAN 2020 competition (Bevendorff et al., 2020) that the problem of misinformation spreaders identification gained the attention of the research community. Most recent studies are focused on analyzing emotional signals (Giachanou et al., 2021), personality and linguistic patterns (Mu and Aletras, 2020; Giachanou et al., 2020). These methods rely on the assumption that the content, and therefore the features that are extracted, remains constant over time. While static linguistic patterns have proven to be useful features for misinformation spreader detection, none of the current methods explore temporal aspects of their behavior. Our model utilizes the users' contextualized content embeddings as user (node) representations and simultaneously leverages their content similarities over time and social interactions dynamically (via edges in the temporal graph).

In the context of user modelling, graph representational learning approaches (Kipf and Welling, 2016; Veličković et al., 2018; Chami et al., 2019) have made significant advances in enhancing NLP models for various tasks (Mishra et al., 2019; Chopra et al., 2020; Sawhney et al., 2021; Kacupaj et al., 2021; Plepi and Flek, 2021). Rath et al. (2020, 2021) identified misinformation spreaders by extracting features from a network that is built based on interpersonal trust metrics. Despite their success, a limitation of the existing approaches is that they do not account for the temporal dynamics of the semantic and social connections.

We argue that the users' characteristics and interactions change dynamically over time due to the dynamic nature of the news cycle, therefore temporal graphs are more suitable to model the evolution of the user-to-user relationships (Wu et al., 2021). Our hypothesis, inspired by Bahns et al. (2017), is that both the social and the content similarity patterns of misinformation spreaders differ from those of other users.

The concept of temporal graphs has been around for some years (Rossi et al., 2020; Seo et al., 2016; Han et al., 2014) with numerous applications (Guo

et al., 2019; Li et al., 2018; Yan et al., 2018). The most relevant to our work is the model proposed by Sawhney et al. (2020), leveraging signals from financial data, social media, and inter-stock relationships via a graph neural network in a hierarchical temporal fashion. We draw inspiration from these approaches and propose a dynamic temporal graph for misinformation spreader detection.

3 Datasets

FACTOID Dataset (Reddit). We utilized the FACTOID dataset published by (Sakketou et al., 2022), which includes a sufficient amount of user history, and, more importantly, simultaneous information on the users' social behavior (Pardo et al., 2020). To the best of our knowledge, this is the only dataset that contains a sufficient amount of social connections to build dense temporal graphs. FACTOID contains a total of 3.3M posts authored by 4.1K users, with 73.8% of the users being "real news spreaders" while the rest 26.2% being misinformation spreaders, determined by the factuality of the news sources they link to. The data covers the period before and after the US elections (from January 2020 to April 2021), making it an ideal dataset for investigating temporal relationships since this time period includes significant events regarding the political scene.

Twitter Dataset. To generalize our content similarity dynamics findings, we utilize in addition the Twitter dataset released by Mu and Aletras (2020). Since the dataset contained the labels and the user IDs, we re-crawled the users' posting history. After filtering the users whose handles were deleted or had insufficient data, the resulting dataset contained 3.5K users and 2.6M posts with roughly 40:60 class distribution of fake and real news spreaders respectively. Since there are practically no social interactions between the users in this dataset, we report results only with the semantic similarity graphs. We split the dataset into train (70%), development (20%) and test (10%) as in the original paper.

	FACTOID	Twitter
Total number of posts	3,354,450	2,626,176
Total number of users	4,150	3,541
# of misinformation spreaders	3,064	1,455
# of real news spreaders	1,086	2,086

Table 1: Summary of dataset statistics for FACTOID and Twitter.

4 Temporal Graph Construction

4.1 Encoding Users

Each user u^i is associated with a posting history $\mathcal{H}^i$. We partition the complete posting time period in equal discrete time frames τ, containing the users' posts that were posted within these time frames.

User2Vec. We adopt User2Vec (Amir et al., 2016) to compute each user's representation $E^i_\tau \in \mathbb{R}^{200}$ based on their corresponding historical posts within the time frame τ, by optimizing the conditional probability of texts given the author.

UBERT. In addition, we use Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) to encode each user's individual historical posts, and we obtain each user's temporal historical encoding $E^i_\tau \in \mathbb{R}^{768}$ by averaging over the posting history length within a corresponding time frame τ.

4.2 Individual Graph construction

We model the user's temporal relationships by constructing a sequence of graphs $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_T$ corresponding to each time frame τ. Each graph $\mathcal{G}_\tau$ is comprised by a set of user nodes $\mathcal{V}_\tau$ that have posted at least once within the time frame τ and a set of edges $\mathcal{E}_\tau$ between these users. We construct the following types of graphs.

Semantic graph. The user embeddings E^i_τ represent each user's context within the time period τ. Users with semantically similar content are close in the vector space (Reimers and Gurevych, 2019) since they have similar context encoding. To construct the users' semantic graphs $\mathcal{G}^{sem}_\tau = (\mathcal{V}_\tau, \mathcal{E}^{sem}_\tau)$, we calculate all the pairwise cosine similarities between the users' embeddings within a time period τ; $cos(E^i_\tau, E^j_\tau)$. We form connections between two users only if their cosine similarity is above a high threshold θ, representing the semantic similarity between two users.

Social graph. On Reddit, users engage in various discussions with their peers. Social science argues that like-minded people tend to interact more with each other (Bahns et al., 2017), therefore, for the FACTOID dataset, we are able to construct the social graph $\mathcal{G}^{soc}_\tau = (\mathcal{V}_\tau, \mathcal{E}^{soc}_\tau)$ in a way that captures the users' social interactions with each other. We define as social interaction the replies and mentions in a post thread. For each thread of posts, we connect all the chain of replies to the root (i.e.

the original post) of the conversation and all mentions/replies to each other. Next, these post connections are translated to user connections in the social graph (Appendix A.2). In the Twitter dataset, the social connections are too few therefore we were unable to build dense temporal graphs.

4.3 Temporal Analysis of Graphs

To answer the *RQ1*, we wish to monitor the temporal evolution of the users' semantic similarities and social interactions between different groups of users over time and associate those temporal fluctuations to the political landscape. We group the users by their credibility label (misinformation spreaders, real news spreaders) and define three different *edge types*: (1) edges between misinformation spreaders ('m2m'), (2) edges between real news spreader ('r2r') and (3) edges between misinformation spreaders and real news spreaders ('m2r'). We partition the users' total posting period (from the start of January 2020 until the end of April 2021) to 16 monthly time periods, and we compute the connections' percentage within each time period for all edge types. The connections' percentage can be interpreted as the normalized edge count of a particular edge type during a time period τ (see Appendix A.4 for more details). For the temporal semantic graphs, an increase in this metric essentially shows an increase in the language usage similarity between different user groups. Correspondingly, for the social graphs, an increase would show that two user groups engage in discourse and share opinions in a thread.

Can we detect different temporal relationship patterns depending on the users' credibility?

Figure 2 depicts the connections' percentage on the semantic graph and the social graph. For both graphs, we can observe that the 'm2r' connections percentage is consistently the lowest for all time periods, indicating that on an aggregate level, misinformation spreaders and real news spreaders do not have as much context similarity to each other and avoid socially interacting with each other. On the other hand, misinformation spreaders seem to be more densely connected with each other and tend to exchange information regularly.

How do the users' temporal semantic and social relationships fluctuate based on the political scene?

Interestingly, we observe peaks in the connections' percentage during January 2020 (event 1), November 2020 (event 2) and January 2021 (event 3) for both graphs. The percentage fluctuations are more obvious in the semantic graph compared to the social graph, this is the first indication that the temporal context similarities might be more useful for the model compared to the social interactions. We provide a list of pivotal political events in Table 2 which evidently explain the increase in the connections' percentage and provide an intuition behind the users' behavior.

Date	Event Description
Feb 5	Trump is acquitted on the charges of abuse of power and obstruction of Congress. (event 1)
Aug 11	Joe Biden chooses Senator Kamala Harris (D-CA) as his running mate
Nov 3	2020 United States elections (event 2)
Jan 6	US Capitol is attacked by supporters of Trump (event 3)

Table 2: **Major political events**[1]. These events are referenced in Figure 2.

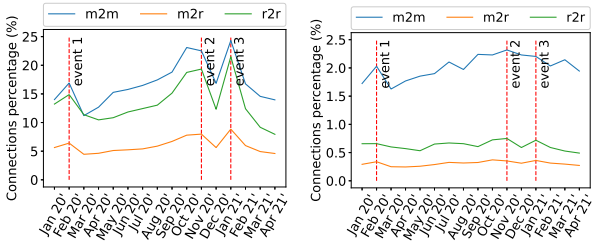


Figure 2: **Connection percentage** of per month for the semantic (left) and social graphs (right). The events shown in this Figure correspond to the events mentioned in Table 2.

5 Neural Network Design

5.1 Graph Neural Network Layer

We utilize three different types of Graph Neural Network (GNN) layers in order to demonstrate the robustness and predictability of the users' connections. The input to the GNN layer is a set of user embeddings E_τ^i for each time frame τ. The GNN layer is shared across the time frames and produces new representations $\widetilde{E}_\tau^i$ which are learned by utilizing either the semantic or social graphs.

Graph Convolutional Neural Network. To embed the nodes in our graph, we employ Graph Convolutional Networks (GCN) (Kipf and Welling, 2016). GCN is a commonly used, powerful graph embedding method that encodes both local graph structure and features of the nodes, by using a layer-wise propagation rule.

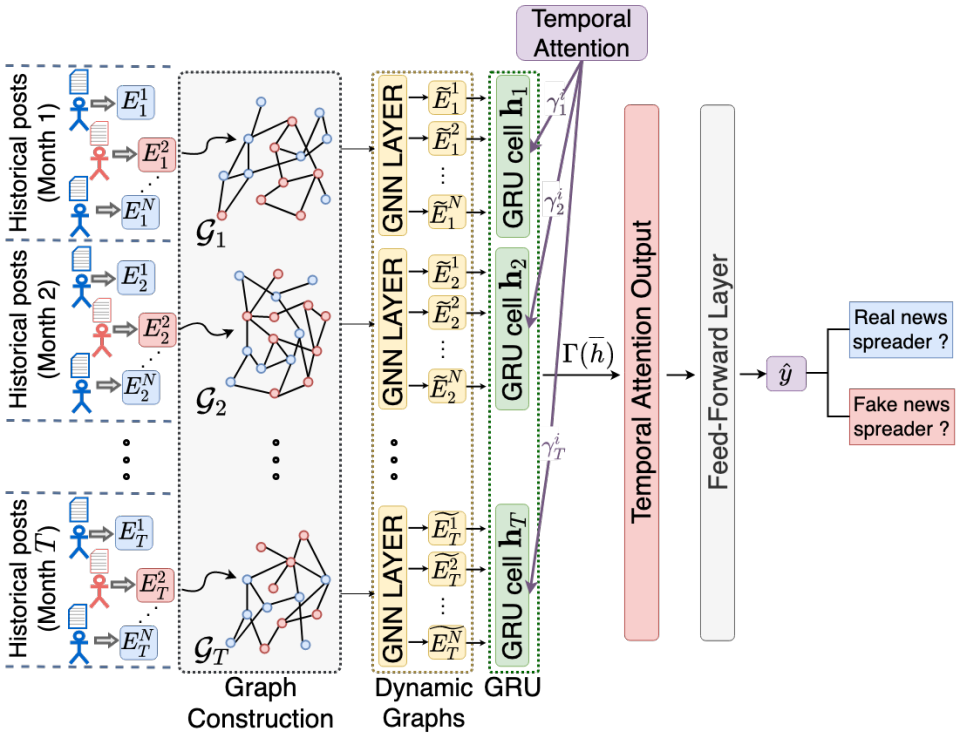


Figure 3: **Overview of the proposed framework.** We first obtain the user embeddings for each time frame and construct the temporal graphs. Next, we feed the graphs to a GNN to extract neighbourhood features. For each user, we use a GRU with temporal attention to compute an overall representation of the user, which is finally forwarded to a classification layer.

Graph Attention Network. As users have a different influence on one another, we need to focus on users that have more relevant connections with higher influence. To model the importance of the influences of the neighbourhood to a node, we use Graph Attention Networks (GAT) (Veličković et al., 2018). GAT attends to the neighborhood of each user and assigns an importance score to the connections that contribute more to the detection of misinformation spreaders.

Hyperbolic Graph Convolutional Neural Networks. Research has shown that GCNs often do not generalize well to hierarchical, tree-like networks such as the social graphs constructed from social media threads (Chen et al., 2012b), since they operate in the Euclidean space. Building on the scale-free nature of the users' social graphs, we utilize Hyperbolic Graph Neural Networks (HGCN) (Chami et al., 2019) which employ graph convolutions in the hyperbolic space as opposed to the standard graph convolutions. The HGCN layer projects the user embeddings in the hyperbolic space to minimize distortions and learn better representations.

5.2 Temporal Neural Network Layer

Temporal Encoding. We investigate the users' behavior over a long time-period, and we wish to encode the dynamic changes between the users' interactions over time. We argue that simply compressing the users' semantic and social connections into one static graph, would introduce too much noise and the information regarding the temporal fluctuations of the semantic and social relationships would be lost. To this end, we model the sequential dependencies through time for each user, with a Gated Recurrent Unit (GRU) (Cho et al., 2014). The GRU encodes the dynamic user graph representations across the time axis, producing hidden states for each time frame τ.

Temporal Attention and Network Optimization. The GRU models the sequential dependencies of the temporal graph user representation, however during the long time span of the users' posting activity, certain socio-political events, such the election seasons, the release date of a new vaccine, etc., may cause the outburst of misinformation spreading. Therefore, we wish to model the contributions of these important time periods to the users' overall

representation. To this end, we employ an attention mechanism (Bahdanau et al., 2016) to compute an overall representation for the user with adaptive weights over the aggregated GRU hidden states.

We formulate the author profiling problem as a binary classification task to predict the class y^i of the user, where $y^i \in \{$misinformation spreader, real news spreader$\}$. The overall learned representations for each user are forwarded into a linear layer, and we use cross-entropy loss to calculate the difference between the true and predicted labels.

6 Experimental Setup

To answer the *RQ2*, we need to investigate the reliability of the temporal semantic and social connections as features for identifying misinformation spreaders in various scenarios.

Predicting future user behavior. We analyze whether the past user behavior, represented through temporal graphs, can be used to predict their future user behavior. To this end, we use the whole set of users in the training, validation and test, but each set contains data from different time periods. Specifically, the training set consists of 8 months (Jan-Aug 20'), and the validation (Sep-Dec 20') and test sets (Jan-Apr 21') 4 months each, resulting in a consecutive 50:25:25 *time split* of the user's posting history. This stands for both datasets since they were collected around the same time period. We provide a visual depiction of this split in Appendix A.5 in Figure 8a.

Generalizing to unseen users. We examine which types of relationships have the ability to generalize to unseen users. In this setup we utilize a *user split*, where we divide the users into a train:validation:test sets of ratio 70:10:20 using all of their posting history. This split is also visually depicted in Appendix A.5 in Figure 8b.

Performance on unseen users in the future. We also aim to test whether the temporal graph features generalize on both unseen users and future content, to this end we utilize the *mixed split*. We split the users into a train:validation:test sets of ratio 70:10:20, where the train set contains users who have posted the first half (Jan-Aug 20') of the whole time period, while the validation and test sets contain a different set of users who post on the second half (Sep 20'-Apr 21'). With this setup, we evidently demonstrate the reliability of the proposed model of detecting misinformation spreaders on unseen data. A visual depiction of this split is provided in Appendix A.5 in Figure 8c.

7 Experimental Results

7.1 Performance results

Feature baselines First, we compare the proposed model to simple, yet strong content-based baselines by utilizing interpretable classifiers; Support Vector Machines (SVM), Logistic Regression (LR), and Random Forest (RF) using the following features:

ngrams: While word ngrams are considered as simple features, they have been used successfully in the past for identifying misinformation spreaders (Vogel and Meghana, 2020). In this case, we utilized the word bi-grams.

statistical-emotional (StEm): We employ a feature vector ($n = 22$) with standard statistical linguistic variables (such as min, max, average number of tokens and characters, lexical diversity, etc.) (Buda and Bolonyai, 2020; Pardo et al., 2020). Additionally, we added 8 emotional dimensions to this baseline feature (Fersini et al., 2020; Mohammad and Turney, 2013).

UBERT: We use the SBERT embeddings of the documents averaged over the whole time frame as feature vectors.

U2V: We also utilized the User2Vec embeddings to represent the users as feature vectors.

Table 3 shows the accuracy results of the baseline models compared to the dynamic graph models on the FACTOID and Twitter datasets. Note that we utilized both the social and the semantic graph and two initialization methods for the FACTOID dataset - in this table we report the best performing variant (for all variants see Table 4). For the Twitter dataset, we experiment only with the semantic graph since there are no social connections between users, and we obtained the temporal graphs with UBERT. We observe that all the proposed models significantly outperform all baseline models for both datasets. For the FACTOID dataset, the best performing dynamic graph model showed higher macro F_1-score compared to the baseline models in all splits, which was on average 10.47% higher on the time split, 15.3% on the user split and 14.08% on the mixed split. For the Twitter dataset, the best performing dynamic graph model showed on average 8% better performance on the time split, 10.8% on the user split and 16.8% on the mixed split.

The results on both datasets validate our claim

	FACTOID									Twitter								
	Time Split			User Split			Mixed Split			Time Split			User Split			Mixed Split		
	SVM	LR	RF	SVM	LR	RF	SVM	LR	RF	SVM	LR	RF	SVM	LR	RF	SVM	LR	RF
ngrams	43.6	56.4	55.4	43.4	58.4	59.5	42.5	42.5	57.6	73.9	75.2	76.9	61.7	65.5	66.6	52.37	42.6	64.81
StEm	52.5	51.6	56.8	49.1	54.9	60.6	54.1	52.1	60.3	61.4	60.8	70.2	59.4	57.3	63.9	43.0	43.5	63.6
UBERT	42.5	47.9	56.1	53.9	58.6	49.7	42.3	45.7	54	62.6	77.3	71.9	64.1	64.7	64.3	36.2	59.4	65.8
U2V	47.6	52.1	61.3	50.2	55.1	56.5	46.4	53.0	59.6	-	-	-	-	-	-	-	-	-
DyGAT	**64.56***			63.59			63.22			**78.2***			67.30			**69.2***		
DyGCN	64.18			65.75			**64.23***			66.9			65.60			66.1		
DyHGCN	64.24			**66.75***			58.58			67.7			**73.90***			65.3		

Table 3: **Baseline experimental results on the FACTOID and Twitter datasets.** Bold indicates the best macro F_1-score. All results are in percentages. We show that the DyGNN framework outperforms all baselines for each split in both datasets. The results with the asterisk (*) are statistically significant based on the Wilcoxon signed rank test ($p = 0.001$) compared to all the baseline methods.

		Semantic			Social		
		Time	User	Mixed	Time	User	Mixed
UBERT	DyGAT	**64.56***	57.26	60.46	62.91	61.66	63.12
	DyGCN	63.57	58.67	61.60	64.18	61.08	59.44
	DyHGCN	55.39	**66.75**	55.25	56.38	62.02	58.58
U2V	DyGAT	63.03	63.59	62.88	63.50	63.01	**63.22***
	DyGCN	62.28	65.75	**64.23***	62.76	64.21	61.35
	DyHGCN	42.51	42.52	47.39	**64.24***	**66.09***	56.10

Table 4: **Comparative analysis of two embedding methods** for semantic graph construction and DyGNN initialization (social graph). Reported macro F_1-score for the FACTOID dataset. All results are in percentages. Bold indicates best result. The results with the asterisk (*) are statistically significant based on the Wilcoxon signed rank test ($p = 0.001$) compared to the second best performing method.

that the specific language features become quickly outdated, while temporal semantic similarities and social interactions are more robust and constitute a better tool for (a) predicting future behavior (time split), (b) predicting the behavior of unseen users (user split), and (c) identifying misinformation spreaders on unseen data (mixed split).

Comparison of dynamic graph models. Table 4 shows the performance results on the three different experimental setups (see Appendix A.6.1 for more detailed results). We analyze the results of the dynamic graph models, based on the utilized graph type (semantic and social), initialization method (UBERT and User2Vec) and graph neural network type (GAT, GCN and HGCN).

Comparing graph types. We observe that the model obtains a slightly better performance by utilizing the semantic similarity graphs compared to utilizing the social graphs for all three setups. Figure 2 shows that the percentage of temporal connections is higher, and fluctuates more, on the semantic graphs compared to the social graphs. This may represent users sharing similar opinionated news regarding the same event, with patterns changing for a new event, while social connections stay similar.

Comparing initialization methods. When UBERT and User2Vec are used in the social graphs, they simply act as initialization vectors, since the social graph construction does not depend on the embedding method. When the models use the social graphs, User2Vec initialization produces better results than UBERT in all setups, despite its lower dimensionality. This performance is expected since User2Vec yields better results than UBERT when it is utilized as a baseline method (Table 3).

The semantic similarity graphs, on the other hand, differ when constructed with UBERT or with User2Vec. In the time split evaluation setup, the semantic graph model achieves the best performance with UBERT, while in the mixed split, the best performance is obtained with User2Vec. This is likely due to UBERT particular suitability for capturing meaningful user similarities even with a small amount of user history, since SBERT (from which we obtain UBERT) is tailored for producing sentence embeddings comparable using cosine-similarity. User2Vec requires a significant amount

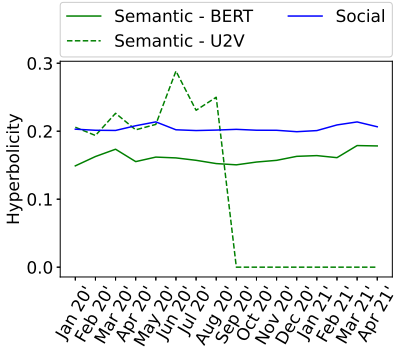


Figure 4: Average hyperbolicity per month

	Semantic			Social		
	Time	**User**	**Mixed**	**Time**	**User**	**Mixed**
DyGNN	**64.56***	66.75	**64.23***	**64.24***	**66.09***	**63.22***
no temporal	55.14	53.53	60.24	62.64	59.37	56.54
no attention	62.27	**66.78***	61.97	61.01	64.51	56.32

Table 5: **Ablation study - temporal dynamics.** In this study we remove the temporal component (keeping simple "static" GNN approach) and the attention. Results show that both components play a significant role to the model's performance. Bold indicates the best macro F_1-score. All results are in percentages. The results with the asterisk (*) are statistically significant based on the Wilcoxon signed rank test ($p = 0.001$).

of documents in order to obtain high-quality user representations however, it leads to a stronger generalizability on unseen data.

Comparing dynamic graph neural networks. We observe that the hyperbolic DyHGCN obtains the best performing results in 3/6 combinations of split and graph type. However, it performs poorly when it utilizes the User2Vec semantic graphs. Figure 4 shows the average hyperbolicity of the dynamic graphs for each month. As is known, high hyperbolicity values indicate a tree-like structure of the network Chen et al. (2012a); Aparicio et al. (2015). Due to the lower posting activity during the last months, and thus higher sparsity of the topics represented by one user, users are more dissimilar, resulting in fewer edges. This in turn leads to lower hyperbolicity during this time period, which explains the DyHGCN's poor performance with User2Vec semantic graphs. The social graph shows high hyperbolicity for all months, therefore DyHCGN achieves superior performance when utilizing the social graphs. DyGAT and DyGCN obtain the best performance once, but in contrast to DyHGCN, they both achieve results within a certain range which is neither too low nor too high.

Discussion. In conclusion, based this comparative analysis, dynamic semantic similarity graphs lead to better results than dynamic social graphs, and given a large amount of user history, User2Vec is preferred for constructing these. In addition, the use of DyHGCN is recommended only when the hyperbolicity of the graph is high, alternatively, DyGAT or DyGCN provide comparable results.

7.2 Ablation Study - Temporal Components

We perform an ablation study on the components of the best performing dynamic graph model to demonstrate the effect of each layer on the overall performance, namely the temporal attention and the temporal graphs:

No attention. We remove the temporal attention layer from our dynamic graph model. Intuitively, this component should focus on the time periods with high misinformation spreading activity and highest differences between user groups.

No temporal dynamics. We average each user's representations across all time frames to obtain a single user representation, and remove the dynamic part of our model by merging all the graphs constructed for every discrete time frame. Specifically, we construct a single graph that includes all the user connections from all time periods and replace the GRU layer, with a linear layer. This model captures the overall semantic and social interactions of the users over their whole posting timeline, and could also be considered as a graph-based baseline.

Table 5 shows the ablative results over the components of the best performing dynamic graph models for all setups. We observe that removing the temporal information has a significant detrimental effect on the performance in all cases, which is on average 7.53%. This demonstrates the strong predictive power of temporal patterns in semantic and social relationships for identifying misinformation spreaders and validates our proposed framework for dynamically modeling the users' semantic and social graphs. In addition, except for the semantic graph on the user split, adding the temporal attention over the users' timeline increases significantly the performance, reinforcing our hypothesis that the similarity of language use during important socio-political events is strongly indicative of misinformation spreading. We have seen that for the semantic graph using the user split, the attention weights through different time slots are the same. Due to this reason, the overall user representation is just a simple average of the GRU states. One reason why this is happening, is because the temporal attention is not capturing temporal patterns of the users, that can generalize to unseen ones.

7.3 Error Analysis

We conducted an analysis of users that consistently get the same prediction *by at least half* of the GNN models. We identify two groups of users; consistently correctly classified, and consistently misclassified. The following error analysis is based on the results obtained on the FACTOID dataset on the user split, however similar results were observed for the rest of the splits.

Approximately 72% of the consistently misclas-

sified users are misinformation spreaders, which can be attributed to the class imbalance decreasing the recall.

It is harder to identify users that are borderline fake news spreaders. Table 6 shows, for the correctly classified and misclassified fake news (FNS) and real news spreaders (RNS), the average number of fake and real news posts, average science and factual level provided in Sakketou et al. (2022) and the average no. of months of active posting. The science level of each user $\in [-1, 1]$ is the normalized weighted average of non-scientific (-1) and scientific (1) articles and the factual level $\in [-3, 3]$ is the normalized weighted average factuality of the news domains, manually labeled by journalists from very low (-3) to very high (3). [2]

		fake posts	real posts	science level	factual level	activity (months)
correctly classified	**FNS**	9.66	39.45	0.13	0.59	12.99
	RNS	0.29	9.95	0.70	1.76	12.57
mis-classified	**FNS**	3.76	22.88	0.16	0.83	11.21
	RNS	0.60	22.67	0.42	1.59	12.37

Table 6: **Error analysis.** Correctly classified fake news spreaders (FNS) post more often than misclassified ones, and post more consistently over time.

As we can see, the misclassified FNS have posted a considerably lower number of fake news on average compared to the correctly classified FNS. While they also posted a lower number of real news posts, their (annotated) factual level is quite high - the source quality plays a role. For the correctly classified FNS, high number of real news combined with low factual level indicates that the real news sources these users are posting are borderline credible - their credibility level is only 'mostly factual'(+1), whereas the credibility level of the fake news sources is from 'low'(-2) to 'very low'(-3). The correctly classified RNS tend to post significantly more scientific articles and articles with higher factuality on average than the misclassified RNS. Overall, correctly classified users of both classes post more consistently over the months compared to the misclassified users.

Since our data heuristics might include wrongly labeled posts and, by extension, users, we manually labeled 210 posts of consistently misclassified users. In this small sample we found that approximately 14% of the posts were wrongly labeled, however less than 1% of the users would obtain a different label because of these posts. We show two examples of mislabeled posts in Table 7.

[2]When embedding the users, we erased the URLs from the text, so that no information about the number of links, or the names of the domains was leaked in the user embeddings, therefore none of the models could have had any prior knowledge of these factors.

Mislabeled as fake news
(...) These pieces rely on discredited sources who have peddled debunked theories about Dominion's supposed ties to Venezuela (...) These statements are completely false and have no basis in fact. (...) [link to non-credible source posting fake news]
Mislabeled as real news
The CCP (Chinese Communist Party) controls Google from within. Change my mind. [link to credible source posting real news]

Table 7: Mislabeled news posts.

8 Conclusion

In this study we proposed a dynamic graph neural network framework that generates temporal graph representations from the users' semantic similarities and social interactions through time.

Our extensive experiments and ablation study demonstrated that the temporal graphs are more efficient than content-based models or simple static graphs for predicting (a) the future misinformation spreading behavior, (b) the behavior of unseen users, and (c) misinformation spreading behavior in a zero-shot scenario. These results indicate that a model utilizing temporal user relationships is more robust and more efficient for misinformation spreader detection compared to topic-sensitive or time-agnostic models, e.g. talking about Trump doesn't make one a misinformation spreader and it is quite normal near election time.

Through exploratory experiments, we analyzed the various aspects of the framework in order to provide an insight into its usability. These experiments showed that dynamic semantic similarities lead to better results than the social ones. The ablation study on the components of the model revealed that the temporal modelling of the users' semantic similarities and social interactions significantly contributes to identifying misinformation spreaders effectively. Our error analysis indicated that the misclassified fake news spreaders tend to post a very low number of fake news posts and a high number of real news posts from highly credible sources. Yet, the proposed framework is applicable as a human moderator-assistance tool for identifying users that post fake news more consistently.

Acknowledgements

This work has been supported by the German Federal Ministry of Education and Research (BMBF) as a part of the Junior AI Scientists program under the reference 01-S20060.

Ethical Considerations and Limitations

Ethical considerations. The ability to automatically approximate personal characteristics of online users in order to improve natural language classification algorithms requires us to consider a range of ethical concerns. Use of any user data for personalization shall be transparent, and limited to the given purpose (Hewson and Buchanan, 2013). Any user-augmented classification efforts risk invoking stereotyping and essentialism, as the algorithm labels people as misinformation spreaders or not. Such stereotypes can cause harm even if they are accurate on average differences (Rudman and Glick, 2012). These can be emphasized by the semblance of objectivity created by the use of an algorithm (Koolen and van Cranenburgh, 2017).

We acknowledge that our research could be used in order to identify gullible individuals that are susceptible to fake news, which enables malicious parties to promote their propaganda. However, the intended use of this research is to limit the misinformation spread by addressing this problem at its origin, therefore our data and the code implementation provided in this work, should only be used for research purposes.

Other limitations. Automatically labelled datasets should be utilized with caution since they might include wrongly labeled posts and, by extension, wrongly labeled users. For example, a number of posts contained multiple links from mixed sources (credible and non-credible). In this paper, we utilized the same labeling method of such posts as Sakketou et al. (2022), where a post is considered misinformation when there is at least one non-credible news source cited. This includes cases where the number of real news sources overcomes the number of fake news sources within one post. We argue that the ratio of the non-credible to credible news sources posted in one post should be considered as a labeling threshold instead. More specifically, if more than half the sources within one post are non-credible, only then should it be labeled as misinformation.

We acknowledge that there is a very thin line separating real news spreaders and misinformation spreaders, however in future works a new class of "potential misinformation spreaders" could be introduced for the users that are on the fence.

References

Silvio Amir, Byron C. Wallace, Hao Lyu, Paula Carvalho, and Mário J. Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 167–177, Berlin, Germany. Association for Computational Linguistics.

Sofía Aparicio, Javier Villazón-Terrazas, and Gonzalo Álvarez. 2015. A model for scale-free networks: Application to twitter. *Entropy*, 17(8):5848–5867.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Angela Bahns, Chris Crandall, Omri Gillath, and Kristopher Preacher. 2017. Similarity in relationships as niche construction: Choice, stability, and influence within dyads in a free choice environment. *Journal of Personality and Social Psychology*, 11:329–355.

Janek Bevendorff, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, Günther Specht, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, and Eva Zangerle. 2020. Overview of pan 2020: Authorship verification, celebrity profiling, profiling fake news spreaders on twitter, and style change detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 372–383, Cham. Springer International Publishing.

Jakab Buda and Flora Bolonyai. 2020. An Ensemble Model Using N-grams and Statistical Features to Identify Fake News Spreaders on Twitter–Notebook for PAN at CLEF 2020. In *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. 2019. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems*, 32:4868–4879.

Wei Chen, Wenjie Fang, Guangda Hu, and Michael W. Mahoney. 2012a. On the hyperbolicity of small-world and tree-like random graphs.

Wei Chen, Wenjie Fang, Guangda Hu, and Michael W. Mahoney. 2012b. On the hyperbolicity of small-world networks and tree-like graphs. *CoRR*, abs/1201.1717.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Shivang Chopra, Ramit Sawhney, Puneet Mathur, and Rajiv Ratn Shah. 2020. Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 386–393. AAAI Press.

E. Fersini, Justin Armanini, and Michael D'Intorni. 2020. Profiling fake news spreaders: Stylometry, personality, emotions and embeddings. In *CLEF*.

Anastasia Giachanou, Esteban A. Ríssola, Bilal Ghanem, Fabio Crestani, and Paolo Rosso. 2020. The role of personality and linguistic patterns in discriminating between fake news spreaders and fact checkers. In *Natural Language Processing and Information Systems*, pages 181–192, Cham. Springer International Publishing.

Anastasia Giachanou, Paolo Rosso, and Fabio Crestani. 2021. The impact of emotional signals on credibility assessment. *Journal of the Association for Information Science and Technology*.

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):922–929.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

Wentao Han, Youshan Miao, Kaiwei Li, Ming Wu, Fan Yang, Lidong Zhou, Vijayan Prabhakaran, Wenguang Chen, and Enhong Chen. 2014. Chronos: A graph engine for temporal graph analysis. In *Proceedings of the Ninth European Conference on Computer Systems*, EuroSys '14, New York, NY, USA. Association for Computing Machinery.

Claire Hewson and Tom Buchanan. 2013. Ethics guidelines for internet-mediated research. The British Psychological Society.

Benjamin D. Horne, Jeppe Nørregaard, and Sibel Adali. 2019. Robust fake news detection over time and attack. *ACM Trans. Intell. Syst. Technol.*, 11(1).

Endri Kacupaj, Joan Plepi, Kuldeep Singh, Harsh Thakkar, Jens Lehmann, and Maria Maleshkova. 2021. Conversational question answering over knowledge graphs with transformer and graph attention networks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 850–862, Online. Association for Computational Linguistics.

Gihwan Kim and Youngjoong Ko. 2021. Graph-based fake news detection using a summarization technique. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3276–3280, Online. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam, a method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, volume 1412.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Corina Koolen and Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia, Spain. Association for Computational Linguistics.

Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. 2017. Rumor detection over varying time windows. *PLOS ONE*, 12(1):1–19.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2018. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting.

Spencer McKay and Chris Tenove. 2021. Disinformation as a threat to deliberative democracy. *Political Research Quarterly*, 74(3):703–717.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Abusive language detection with graph convolutional networks. *CoRR*, abs/1904.04073.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Yida Mu and Nikolaos Aletras. 2020. Identifying twitter users who repost unreliable news sources with linguistic information. *PeerJ Comput. Sci.*, 6:e325.

Francisco M. Rangel Pardo, Anastasia Giachanou, Bilal Ghanem, and Paolo Rosso. 2020. Overview of the 8th author profiling task at PAN 2020: Profiling fake news spreaders on twitter. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Joan Plepi and Lucie Flek. 2021. Perceived and intended sarcasm detection with graph attention networks. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4746–4753, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bhavtosh Rath, Xavier Morales, and Jaideep Srivastava. 2021. SCARLET: explainable attention based graph neural network for fake news spreader prediction. *CoRR*, abs/2102.04627.

Bhavtosh Rath, Aadesh Salecha, and Jaideep Srivastava. 2020. Detecting fake news spreaders in social networks using inductive representation learning.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. 2020. Temporal graph networks for deep learning on dynamic graphs.

Laurie A Rudman and Peter Glick. 2012. *The social psychology of gender: How power and intimacy shape gender relations*. Guilford Press.

Flora Sakketou, Joan Plepi, Riccardo Cervero, Henri-Jacques Geiss, Paolo Rosso, and Lucie Flek. 2022. Factoid: A new dataset for identifying misinformation spreaders and political bias.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. 2020. Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8415–8426, Online. Association for Computational Linguistics.

Ramit Sawhney, Harshit Joshi, Rajiv Ratn Shah, and Lucie Flek. 2021. Suicide ideation detection via social and temporal user representations using hyperbolic learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2176–2190, Online. Association for Computational Linguistics.

Youngjoo Seo, Michaël Defferrard, Pierre Vandergheynst, and Xavier Bresson. 2016. Structured sequence modeling with graph convolutional recurrent networks. *CoRR*, abs/1612.07659.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks.

Inna Vogel and Meghana Meghana. 2020. Fake news spreader detection on twitter using character n-grams. notebook for pan at clef 2020.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2021. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition.

A Appendix

A.1 Dataset

A.1.1 Analysis of the linguistic differences

To get an intuition for the actual linguistic differences between the two user groups of misinformation spreaders and real news spreaders, we extracted the learned token weights from the SVM model in order to study the predictiveness of the tokens for each class (Guyon et al., 2002). The most predictive tokens are shown in Table 8. It can be seen that there's a tendency for misinformation spreaders to reference politically left-leaning groups as "liber", "dem", "left" or "blm" (referring to the Black Lives Matter movement), while real news spreaders use the terms "fascist" and "republican" with higher frequency.

Label	Tokens
Misinformation Spreaders	china, video, come, offici, blm, corrupt, media, away, liber, order, new, trump's, seem, wrong, kill, left, dem, riot
Fact Checkers	public, first, week, understand, trial, fascist, republican, war, one, forced-birth, health, pleas, power, let, shock, view, service

Table 8: Top-ranked tokens for each label.

A.2 Social graph construction

Figure 5 shows the transformation of the thread structure into a social graph.

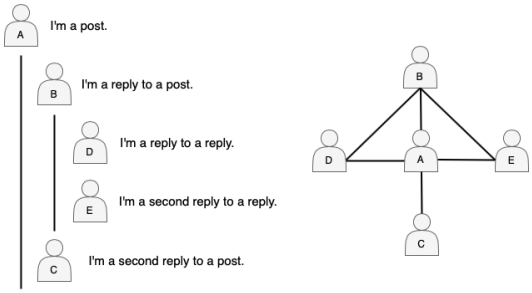


Figure 5: Transforming a post/reply tree in social media into a social graph network.

A.3 Temporal Analysis of Nodes

Centrality. Figure 6 depicts the graph centrality normalized by the number of posts. This metric helps in identifying important nodes in a graph. We can see that, in the linguistic graph, the centrality of the misinformation spreaders and real news spreaders follows a similar pattern but fluctuates a lot over time. Interestingly, there's an obvious increase in the centrality of both classes during August, right

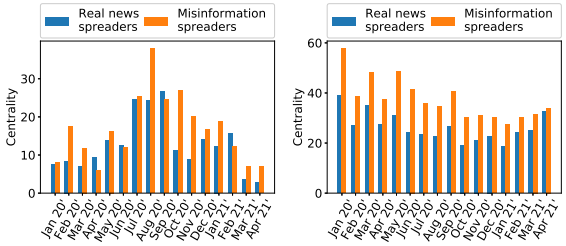


Figure 6: Approximated (k=1000) graph centrality normalized by post amount calculated for all time spans for the semantic (left) and social (right) graph.

after former President Trump announced the possibility of postponing the US elections (see Table 2). This increase is more obvious in the misinformation spreaders, meaning that they are discussing a particular topic more extensively compared to the real news spreaders. In the social graph, we observe a great difference in the values of centrality between misinformation spreaders and real news spreaders. This metric shows that misinformation spreaders are gathered in the center of the graph, while real news spreaders are in the periphery of the graph and are not that densely connected to each other. This essentially indicates that misinformation spreaders form a densely connected "community" and marginalize real news spreaders. The centrality of the misinformation spreaders decreases over time, while in the case of real news spreaders it fluctuates but still stays within a specific range. This apparent dynamically changing behavior of the nodes supports our choice of temporal modelling of the graphs.

Homophily. In Figure 7, we show the amount of homophily observed for both semantic and social graphs, which is defined as the percentage of edges that connect users with the same label. Interestingly, we observe that in the semantic graph the homophily follows different patterns in misinformation spreaders and real news spreaders, and it is fluctuating over time. In the social graph, the misinformation spreaders have consistently higher homophily than real news spreaders, which means that they tend to interact and exchange opinions more with each other compared to real news spreaders. These results complement the edge analysis from Section 4.3 which shows that users from the same credibility group tend to socially interact more with each other, which is more apparent in misinformation spreaders.

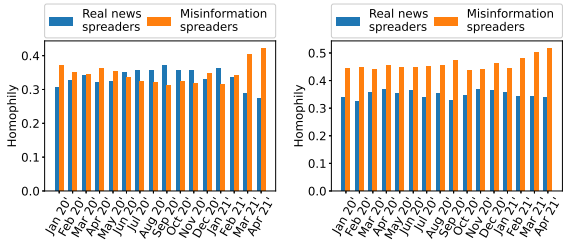
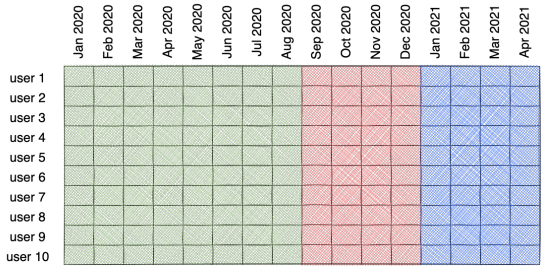
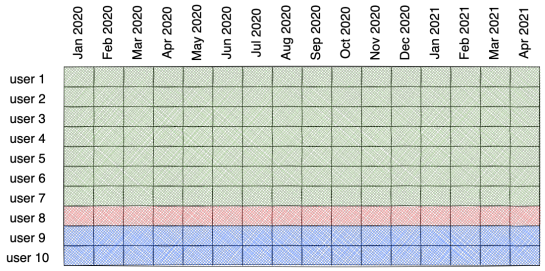


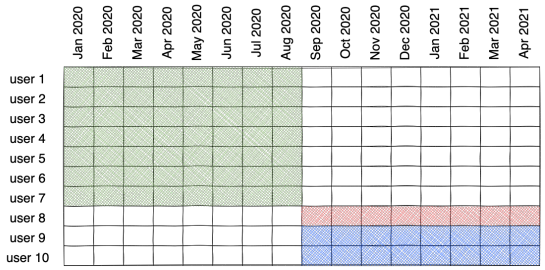
Figure 7: Amount of homophily observed through time for both semantic (left) and social graph (right).



(a) Time split. Splitting the time periods in order to predict future user behavior.



(b) User split. Splitting the users in order to predict the behavior of unseen users.



(c) Mixed split. Splitting the users and the time periods in order to predict the behavior of unseen users in the future.

Figure 8: Visual demonstration of the (a) Time split, (b) User split and (c) Mixed split.

A.4 Connections' percentage

We define the connections' percentage of a certain edge type as $\rho_{\text{edge type}} = r^{(\tau)}_{\text{edge type}} / R^{(\tau)}_{\text{edge type}}$, where $r^{(\tau)}_{\text{edge type}}$ is the number of edges (of that edge type) that exist between two users during the time period τ and $R^{(\tau)}_{\text{edge type}}$ is the number of all possible connections (of that edge type) at the time period τ, computed as follows:

$$R^{(\tau)}_{m2m} = N^{(\tau)}_{m}(N^{(\tau)}_{m} - 1)/2$$
$$R^{(\tau)}_{r2r} = N^{(\tau)}_{r}(N^{(\tau)}_{r} - 1)/2$$
$$R^{(\tau)}_{m2r} = (N^{(\tau)}_{m} + N^{(\tau)}_{r})(N^{(\tau)}_{m} + N^{(\tau)}_{r} - 1)/2$$

where $N^{(\tau)}_{m}$ is the number of misinformation spreaders and $N^{(\tau)}_{r}$ is the number of real news spreaders that have posted at least one post at time period τ.

A.5 Training Setup

We use the pretrained model `'all-mpnet-base-v2'` from SBERT[3], which achieved the best performance on various challenging similarity datasets (Cer et al., 2017). This model has max length set to 512, uses mean pooling and has the output dimension $d_b = 768$. The users' historical representations are obtained as described in Section 4.1 For each post in the user history, we masked the links so that the cosine similarity is not attributed based on the links. We run experiments with $\delta \in 15, 30, 60, 360$ (δ is the number of days spanned by each that each time period τ). In each sample, we randomly sample $n \in 200, 400, 800, 1200$ users, and we build a subgraph of those users for each discrete time window. In the semantic graph, we connect users with each other based on the hyperparameter $\theta \in [0, 1]$ (as defined in Section 4.2). We find out that our model works best with the following hyperparameters: $n = 200$, $\delta = 30$, $\theta = 0.8$. For the models initialized with User2Vec embeddings, we use the dimensions $d_g = 100$ for our graph layer and $d_r = 50$ for our GRU sequential layer. On the other hand, for the models initialized with UBERT embeddings we use the dimensions $d_g = 256$ for our graph layer and $d_r = 128$ for our GRU sequential layer. We use Adam optimizer (Kingma and Ba, 2015) with learning rate $5e-5$, weight decay $1e-2$, and train the model for 100 epochs using early stopping with patience 20 on the validation set. We run each experiment with 5 random seeds and report the mean result on the test set in Tables 3, 4 and 5. DyGAT model using User2Vec embeddings as initialization has 116K parameters, while DyGCN and DyHGCN have 55K parameters. On the other hand, DyGAT

[3] https://www.sbert.net/docs/pretrained_models.html

		Semantic graph								
		Time Split			User Split			Mixed Split		
		F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall
UBERT	DyGAT	**49.64**	**45.09**	**55.22**	33.44	40.46	28.49	43.18	40.09	46.77
	DyGCN	46.55	45.52	47.63	36.8	41.06	33.33	44.44	41.9	47.31
	DyHGCN	45.97	34.3	69.65	**52.45**	**48.2**	**57.53**	44.81	33.88	66.13
U2V	DyGAT	47.85	42.89	54.11	42.86	54.1	35.48	44.44	45.98	43.01
	DyGCN	41.47	49.56	35.65	52.09	45.9	60.22	**49.77**	**44.17**	**56.99**
	DyHGCN	0	0	0	0	0	0	10.38	42.31	5.91

Table 9: Reported F_1-score, Precision and Recall on the fake news spreader class for the FACTOID dataset utilizing the semantic graph. All results are in percentages. Bold indicates the best macro F_1-score on both classes.

		Social graph								
		Time Split			User Split			Mixed Split		
		F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall
UBERT	DyGAT	47.14	43.07	52.05	41.42	46.05	37.63	46.43	44.17	48.92
	DyGCN	44.97	51.28	40.04	39.5	47.37	33.87	39.89	40	39.78
	DyHGCN	47.39	35.25	72.29	51.99	40.18	73.66	32.71	53.01	23.66
U2V	DyGAT	57.24	41.84	90.61	43.24	48.98	38.71	**48.36**	**42.92**	**55.38**
	DyGCN	41.85	51.54	35.23	48.9	44.84	53.76	44.05	41.63	46.77
	DyHGCN	**46.74**	**47.74**	**45.78**	**54.47**	**45.07**	**68.82**	46.21	34.78	68.82

Table 10: Reported F_1-score, Precision and Recall on the fake news spreader class for the FACTOID dataset utilizing the social graph. All results are in percentages. Bold indicates the best macro F_1-score on both classes.

using UBERT embeddings as initialization has 1M parameters, while DyGCN and DyHGCN have 427K parameters. Our experiments for each model take around 1 hour to run on NVIDIA A100-PCIE 40GB GPU. Our implementation, the annotated dataset, and the results are publicly available to facilitate reproducibility and reuse.

A.6 Detailed Experimental Results

A.6.1 Comparison of the graph types

Tables 9 and 10 show the F_1-score, Precision and Recall on the fake news spreader class for the FACTOID dataset utilizing the semantic and social graphs respectively. Given the same combination of setups, i.e different splits, GNN and embedding initialization, we qualitatively compared the results obtained by utilizing the semantic and social graphs. We report the findings regarding the cases with the best macro F_1-scores (in bold).

In the time split, for the DyGAT+UBERT model, we observed that the results are not significantly different when comparing the utilization of semantic and social graphs. In the same split, for the DyHGCN+User2Vec model, we note that 24.99% of the users were classified differently by the semantic and social models, this difference is expected since the difference between the F_1-scores obtained by each graph type is more than 20%. When the semantic graph is utilized, we observe that DyHGCN+User2Vec fails to recognize any of the misinformation speaders, however it achieves an impressively high performance with the social graph. This result is justified due to the low hyperbolicity values of the semantic User2Vec graph as mentioned in Section 7.1.

In the user split, for the DyHGCN+UBERT model, we note that 32.54% of the users were classified differently from the semantic and social models, even though the difference between their macro F_1-scores is only 4%. By utilizing the semantic graph, the model yields to a worse Recall for the fake news spreader class, but higher Recall for the real news spreader class. In the same split, for the DyHGCN+User2Vec model, we note that 39.72% of the users were classified differently, however this difference is expected since the F_1-scores obtained by the semantic and social models have more than 20% difference between them. Once more we observe a staggering difference between the F_1-scores obtained from semantic and social models, with the social model achieving the highest score.

In the mixed split, for the DyGCN+User2Vec

model, we note that 27.55% of the users were classified differently. We observe that the model obtains higher recall on the fake news spreader class when the semantic relationships are utilized, instead of the social ones. In the same split, for the DyHGCN+UBERT model, we observe that 7.82% of the users were calculated differently. By utilizing the social graph, the model achieves higher Recall on the fake news spreader class.