# Temporal Alignment of Reddit Network Embeddings

Siobhán Grayson and Derek Greene

Insight Centre for Data Analytics, University College Dublin, Ireland.
siobhan.grayson@insight-centre.org,
WWW home page: https://graysons.github.io/

## 1 Introduction

Motivated by the concepts and findings being developed for diachronic word embeddings, in this paper, we explore how the application of the same principles can be leveraged to study structural roles from a temporal perspective. In the same way words with a similar meaning will repetitively appear in the same contexts, structural roles in graphs are also defined by the topological company that they keep. However, structurally equivalent roles may or may not occur in close proximity within a graph. Our goal is to map the participants of the popular social media website *Reddit*[1], into an embedding space that best represents the similarity of the structural roles that they occupy and to then measure how their roles change over time.

## 2 Methodology

Our dataset consists of 16 subreddits identified by Hamilton and Zhangs in their work on characterising Reddit communities [2][2] as exhibiting the most "loyal" user features (teams and sports related subreddits) and 13 subreddits identified as having the highest "vagrant" user patterns (see Table 1). When identifying loyal and vagrant communities, Hamilton et al. considered user commenting behaviour on Reddit over time and defined loyal and vagrant users as follows: Loyal members are users who for two consecutive months have submitted at least 50% of their comments to one Subreddit. Vagrant members on the other hand are defined as users who comment 1 to 3 times within a Subreddit in one month but then do not submit any comments the subsequent month despite still being active on Reddit. For temporal analysis, we partitioned data from late January to October 2014 into three windows each consisting of three months.

| **Class** | **# SR** | **# $V_{T1}$** | **# $E_{T1}$** | **# $V_{T2}$** | **# $E_{T2}$** | **# $V_{T3}$** | **# $E_{T3}$** |
|---|---|---|---|---|---|---|---|
| Loyal | 13 | 15,319 | 89,496 | 15,193 | 91,138 | 14,531 | 87,149 |
| Vagrant | 16 | 13,462 | 22,323 | 14,030 | 23,831 | 13,314 | 22,247 |

**Table 1:** Notation - SR: Subreddits, $V_{T1}$: Nodes Temporal Window 1, $E_{T1}$: Edges temporal window 1.

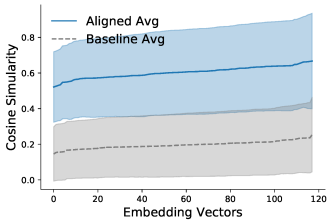
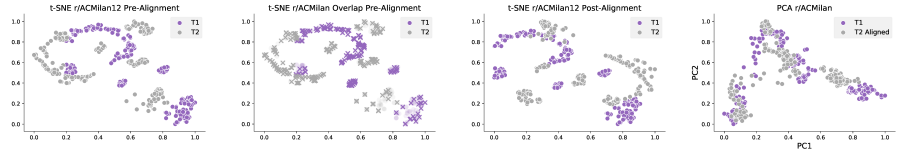



**Fig. 1:** Cosine Similarity for alignment evaluation.

[1]The url address for this site is: https://www.reddit.com/

[2]Further details can be found on the webpage where the dataset is available to download: http://snap.stanford.edu/data/web-RedditNetworks.html

For the purposes of this study, two users are defined as having corresponding roles if their occurrences within the Reddit networks are structurally equivalent. To assess user role variation over time, we first select the 100 highest frequency participants for each 3 months and then use the overlap of this set that spans all window partitions to extract temporally related networks. Once we have our temporal networks, actors are then described in terms of their roles by applying the directed and weighted version of the graph embedding algorithm, *struc2vec* [4], specifically designed to capture structural equivalence between nodes.




(a) Pre-Alignm. Emb. (b) User Overlap Emb. (c) Post-Alignm. Emb. (d) PCA Emb.

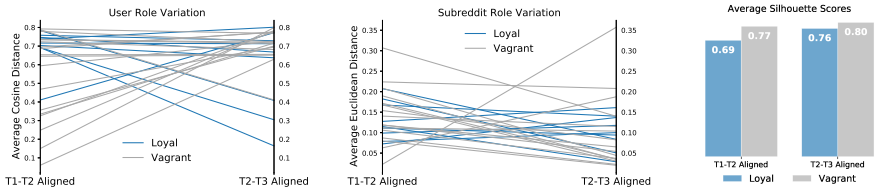
**Fig. 2:** Loyal subreddit 'r/ACMillan' before and after alignment, and then dimension reduction.

The embedding spaces in this study are then aligned using normalised orthogonal Procrustes, an approach popular for aligning diachronic word embeddings [1, 5], as it derives the optimal rotation of a "source" matrix with respect to a "target" matrix without scaling by minimising the sum of squared distances between elements. Alignments can be evaluated by generating a second embedding matrix for the same time period and comparing the cosine similarity between vectors. Fig.1 displays the average of aggregated cosine similarity results ($1/N\sum_{i=1}^{N}\cos(\boldsymbol{v}_i^t,\boldsymbol{v}_i^{t+\Delta})$) and the standard deviations computed across all embedding spaces and their duplicates for both before (Baseline) and after alignment. In all cases, rotations reduced the dissimilarity between temporal user embeddings. Fig.2 illustrates the affect of the alignment process by visualising the embeddings for time period 2 (T2) being aligned to time period 1 (T1) using t-SNE [3]. Occasionally, derived anchors were not dispersed throughout the embedding space which resulted in the sign of the eigenvectors being 'flipped' during PCA. To resolve this, further alignment of roles is applied by changing the signs of equivalent principal components to agree if they do not already.

Once embeddings have been aligned, we can compute the cosine distance between an actors embedding at time $t$ and $t+\Delta$: $1-\cos(\boldsymbol{v}_i^t,\boldsymbol{v}_i^{t+\Delta})$ to detect changes in an individual's role across time. Greater distances indicate a larger deviation in the type of roles a participant occupied during different periods and vice versa. We then aggregate individual results to derive a mean cosine distance score for each subreddit so that comparisons can be made across loyal and vagrant user role fluctuations. In order to observe the variation of community roles over time, we first find the maximum number of clusters present across time periods to be compared by decomposing the 128 dimensional embedding spaces into 2 dimensions using PCA. The Elbow method using Euclidean Kmeans is then applied to determine the number of clusters present. The maximum equal cluster number across two embedding spaces is recorded and 1-Nearest Neighbours is applied to compute the Euclidean distance between the closest aligned centroids. The resulting value provides insight into how much the general roles present within a subreddit community have changed over time. Finally, silhouette scores are also computed for each embedding space to determine whether roles evolve to become more or less acutely defined over time.

## 3 Results

The results of our analysis are depicted in Fig.3. The first figure, Fig.3(a), illustrates the average cosine distances computed for each subreddit mapped from time period T1 to aligned T2. The majority of user cosine distances continue to remain as dissimilar to each other in the second temporal embedding space, time period T2 aligned with time period T3. Hence, our preliminary findings suggest that although individual users of Reddit may change role frequently, the universal community level roles remain relatively static in comparison. The static nature of community roles in comparison to user roles is further examined by calculating the average Silhouette Score for each subreddit. The average Silhouette scores, Fig.3, indicate that it's not an isolated scenario.




(a) User Roles (b) Community Roles (c) Role Variation

**Fig. 3:** The temporal user and community role dynamics observed via three different metrics for comparing similarity: Cosine distance, Euclidean distance, and Silhouette scores.

## 4 Conclusion

In this paper, we applied the role embedding algorithm, *struc2vec* to three consecutive temporal windows of user networks and then aligned the resulting embedding spaces using orthogonal Procrustes. Overall, our findings suggest that while participant roles fluctuate a lot, the ubiquitous community roles present are a lot more static. However, further analysis is required and we hope to extend the current work to explore subreddits such as AskReddits, Debate Reddits, Questions Reddits, where roles are generally quite distinguished to allow for further comparisons to be made.

## References


1. W. L. Hamilton, J. Leskovec, and D. Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1489–1501, 2016.
2. W. L. Hamilton, J. Zhang, C. Danescu-Niculescu-Mizil, D. Jurafsky, and D. Jurafsky. Loyalty in online communities. In *International AAAI Conference on Weblogs and Social Media*, volume 2017, page 540. NIH Public Access, 2017.
3. L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
4. L. F. Ribeiro, P. H. Saverese, and D. R. Figueiredo. Struc2vec: Learning node representations from structural identity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 385–394. ACM Press, Aug 2017.
5. S. L. Smith, D. H. Turban, S. Hamblin, and N. Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. 2016.