Temporal Analysis of Reddit Networks via Role Embeddings

Siobhán Grayson
Insight Centre for Data Analytics
Dublin, Ireland
siobhan.grayson@insight-centre.org

Derek Greene
Insight Centre for Data Analytics
Dublin, Ireland
derek.greene@insight-centre.org

ABSTRACT

Inspired by diachronic word analysis from the field of natural language processing, we propose an approach for uncovering temporal insights regarding user roles from social networks using graph embedding methods. Specifically, we apply the role embedding algorithm, *struc2vec*, to a collection of social networks exhibiting either "loyal" or "vagrant" characteristics derived from the popular online social news aggregation website Reddit. For each subreddit, we extract nine months of data and create network role embeddings on consecutive time windows. We are then able to compare and contrast how user roles change over time by aligning the resulting temporal embeddings spaces. In particular, we analyse temporal role embeddings from an individual and a community-level perspective for both loyal and vagrant communities present on Reddit.

CCS CONCEPTS

• **Networks** → **Online social networks**; • **Information systems** → *Temporal data*; • **Mathematics of computing** → Exploratory data analysis;

KEYWORDS

temporal networks, diachronic role embeddings, reddit, embedding alignment

ACM Reference Format:
Siobhán Grayson and Derek Greene. 2018. Temporal Analysis of Reddit Networks via Role Embeddings . In *Proceedings of ACM KDD conference (MLG'18).* ACM, New York, NY, USA, Article 4, 5 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Embeddings are now a common component of the typical text analysis pipeline thanks to their ability to apply vectorially the logic of "You shall know a word by the company it keeps" (Firth, J. R. 1957) and the accessibility of Mikolov's *word2vec* [14]. Not only have word embeddings enhanced translation tasks [25] but they have also exposed the cultural biases that are emeshed within languages [1, 5]. Embeddings have even been extended to study diachronic language characteristics [11]. Diachronic embeddings are created by embedding separate time windows and then aligning the resulting spaces orthogonally such that distances are not warped. This allows for direct measurements to be taken of how much the meaning of semantically similar words change over time in comparison to eachother using distance metrics such as cosine distance [8].

Therefore, 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 repeatability 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 [4]. Hence, by embedding networks into different dimensions the distances between similar entities can be reduced. Our goal is to then map the participants of the popular social media website *Reddit*[1], into an embedding space that best represents the structural roles that they occupy and to then measure how their roles change over time. In particular, we analysis how roles evolve from both an individual and community level perspective. Our findings suggest that while participant roles fluctuate a lot, the ubiquitous community roles present are relatively static in comparison.

2 RELATED WORK

The social news aggregation website Reddit was founded in 2005 and has grown over the years to now have over 200 million unique users. Fives years of Reddit's lifetime has been analysed by Singer et al. [20] from both a user submissions perspective, to how community level attention evolves over time, resulting in "an ever-increasing diversification of topics accompanied by a simultaneous concentration towards a few selected domains". In 2016, Newell et al. conducted a study on user migration in social networks and found that an important factor in Reddit's ability to retain users was the availability of "niche" content not provided elsewhere [15]. Thus, this retention of users suggests a certain level of loyal activity which is explored in further detail by Hamilton and Zhang et al. in 2017 who categorise different subreddits as either "loyal" or "vagrant" communities based on the network and textual characteristics [9]. The vagrant nature of Reddit is also investigated by Leavitt in 2015 via "throwaway accounts" [12], potentially one of the shortest temporal roles present on Reddit.

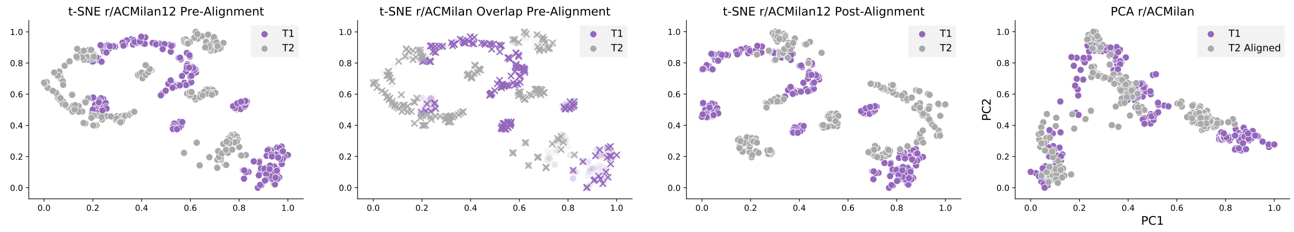
2.1 Graph Embeddings

Increasingly, network scientists are adpoting embedding techniques of their own to examine graphs in different dimensions. Grover et al. even gestering a nod to *word2vec* with their graph embedding algorithm entitled *node2vec* [6]. In short, a diverse range of embedding approaches exist ranging from granular node and edge generated emebeddings [3, 23] to whole graph [16] and dynamic [24]. In fact, there exists at least three comprehensive survey papers

MLG'18, August 2018, London, England UK

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

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



(a) Pre-Alignment Embeddings **(b) User Overlap Embeddings** **(c) Post-Alignment Embeddings** **(d) PCA Embeddings**

Figure 1: Visualisation of the loyal subreddit 'r/ACMillan' before and after alignment and dimension reduction.

on the subject [2, 17, 22]. Reddit is even amoung one of the datasets used by Hamilton et al. to evaluate their attributed graph embedding algorithm *GraphSAGE* [7]. As such, network scientists have a large pool of techniques to select from, each catering for different types of graphs. Although dynamic behaviour has been modelled before [19], dynamic embedding algorithms specifically designed for finding structural equivalences that also incorporate direction and weight are few and far between. For the purposes of this study, three role embedding techniques stood out, Rolx [10], GraphWave [4], and struc2vec [18]. After testing each, we decided to remain with struc2vec as it could be optimised such that computations completed faster than it peers and it generalised well, grouping similar roles together as opposed to over-fitting and identifying them as completely different to each other.

3 METHODOLOGY

In parallel to our motivation to observe whether temporal role variations occur at user and subreddit levels, we're also keen to learn whether role variation is also related to subreddit type. Therefore, we're using a subset of the directed Reddit chain-based interaction networks where users are linked if they comment within a linear chain originally curated by Hamilton and Zhangs in their work on characterising Reddit loyalty[9][2]. Our dataset consists of 16 subreddits identified by Hamilton et al. as exhibiting the most "loyal" user features (teams and sports related subreddits) and 13 subreddits identified as having the highest "vagrant" user patterns. When identifying loyal and vagrant communities, Hamilton et al. considered user commenting behaviour on Reddit over time. They define 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. In doing so, they exhibit a preference and commitment to this 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 9 consecutive months of data, spanning from late January to October in 2014, into three temporal windows consisting each consisting of three months. A summary of this data is provided in Table 1.

Table 1: Summary of Reddit Data

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

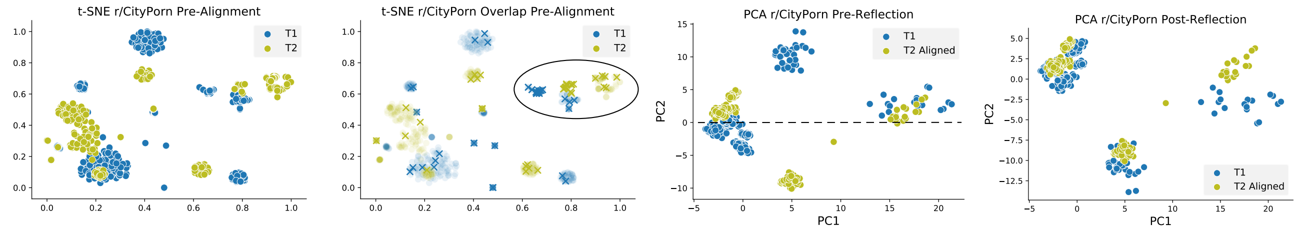
Notation - SR: Subreddits, V$_{T1}$: Nodes in Temporal Window 1, E$_{T1}$: Edges in temporal window 1.

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* [18], specifically designed to capture structural equivalence between nodes. Specifically, similar roles are mapped closer together in the resulting embedding space while dissimilar roles will be further a part. Hence, role similarity can be assessed by measuring the distance between role embeddings. However, before embeddings generated from different time windows can be compared they must first be aligned.

3.1 Temporal Role Alignment

The embedding spaces in this study are aligned using normalised orthogonal Procrustes, an approach popular for aligning diachronic word embeddings [8, 21], 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. This results in the ability to directly compare temporal embedding spaces to each other using dimension appropriate distance metrics. In particular, the orthogonal Procrustes rotation between spaces is computed by mapping the overlapping sets of users to each other. Fig.1 illustrates the process by visualising T2 (time period 2) embeddings being aligned to T1 (time period 1) embeddings using t-SNE [13]. Alignments can then be evaluated by generating a second embedding matrix for the same time period and comparing the cosine similarity between vectors. Fig.3 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.

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



(a) Pre-Alignment Embeddings **(b) User Overlap Embeddings** **(c) PCA Pre-Alignment** **(d) PCA Post-Alignment**

Figure 2: Visualisation of the vagrant subreddit 'r/CityPorn' before and after alignment and dimension reduction.

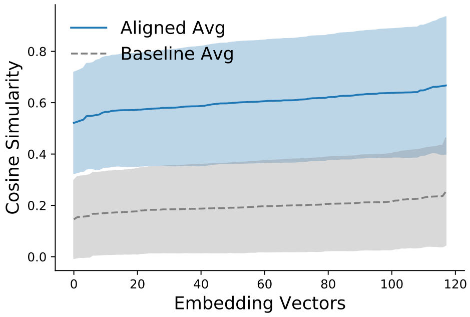


Figure 3: Cosine Similarity results for alignment evaluation.

However, although orthogonal Procrustes appeared to perform reasonably well when evaluated using overlapping user embeddings, it did not always correctly align community role embeddings, defined as the agglomerates of similar roles that emerge after the application of Principal Component Analysis (PCA) to aligned spaces. Specifically, the T1-T2 Aligned and T2-T3 Aligned spaces, where T1-T2 Aligned stands for the embedding space that occurs from the amalgamation of T1 with T2 embeddings that have been aligned to T1, and similarly for T2-T3 Aligned. Unlike our previous example of *ACMilan* embeddings (Fig.1), the majority of overlapping individuals in *CityPorn*'s embeddings are confined to a relative small region of the space (circled in Fig.2(b)). This results in the sign of eigenvectors being 'flipped' during PCA and hence, rather than similar community roles overlapping, they become mirrored in the resulting PCA space (Fig.2(c). To resolve this, further alignment of roles is applied by changing the signs of equivalent principal components to agree if they do not already.

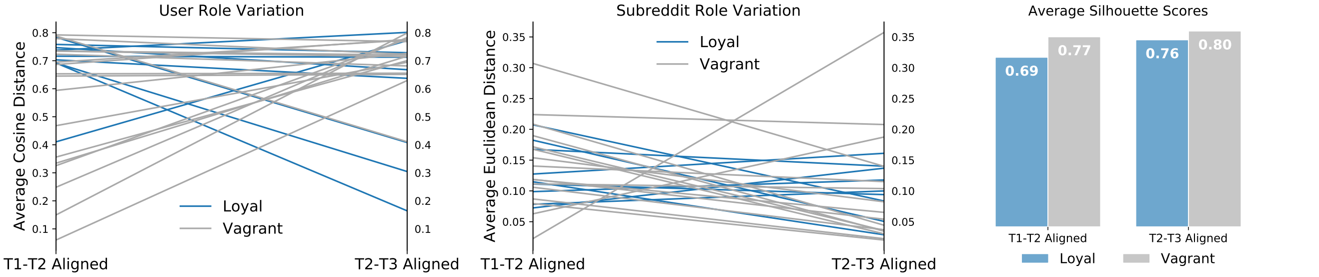
3.2 Measuring Role Variation Across Time

Once embeddings have been aligned, temporal comparisons can be made directly using appropriate distance metrics. To detect changes in an individual's role across time, we compute the cosine distance between an actors embedding at time t and $t + \Delta$: $1 - \cos(\boldsymbol{v}_i^t, \boldsymbol{v}_i^{t+\Delta})$. Greater distances indicate a larger deviation in the type of roles a participant occupied during different periods. While small cosine distances suggest an individual's role has not changed much over time as they map into a relatively similar space. 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 k-Nearest Neighbours, where k=1, 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.

4 RESULTS

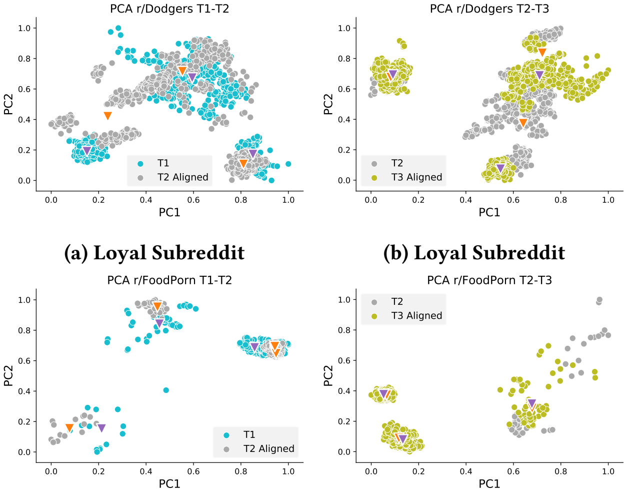
The results of our analysis are depicted in Fig.4. The first figure, Fig.4(a), illustrates the average cosine distances computed for each subreddit mapped from time period T1 to aligned T2. The majority of user cosine distances continue to remain as dissimilar to each other in the second temporal embedding space, time period T2 aligned with time period T3. Although differences can be observed between loyal and vagrant users, such as vagrant users appearing to change roles to a greater extent than loyal users, while loyal users appear to retain the same role over time. It is hard to define this as a general rule that could be applied to all of Reddit's community without applying our investigation to a larger number of subreddits. However, our preliminary findings suggest that although individual users of Reddit may change role frequently, the universal community level roles remain relatively static in comparison. Fig.4(b) depicts how distances between role cluster centroids for both loyal and vagrant subreddits remain small, indicating they are similar to each other.

The static nature of community roles in comparison to user roles is further examined by visualising the PCA projections and by calculating the average Silhouette Score for each subreddit. Visual comparisons of one loyal subreddit (Fig.5(a)(b)), r/Dodgers, and one vagrant subreddit similar in size (Fig.5(c)(d)), r/FoodPorn, depict striking differences in the definition of clusters. The loyal subreddit role clusters are more dispersed in comparison to it's vagrant counterpart where clusters are spread and tightly compact. The average Silhouette scores, Fig.4, indicate that it's not an isolated scenario. However, the differences are small, and again, further subreddits



(a) User roles are not static. **(b) Community roles are relatively static.** **(c) Variation of defined roles.**

Figure 4: The temporal user and community role dynamics observed via three different metrics for comparing similarity: Cosine distance, Euclidean distance, and Silhouette scores.



(a) Loyal Subreddit **(b) Loyal Subreddit**

(c) Vagrant Subreddit **(d) Vagrant Subreddit**

Figure 5: Subreddit PCA projections across time.

will need to be examined before we can say definitively that such differences are significant.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we have analysed 29 subreddits classed as either "loyal" or "vagrant" using a methodology inspired by the study of diachronic word embeddings in the field of natural language processing. Specifically, 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. We found that in certain community role cases, orthogonal Procrustes was not enough to align spaces entirely if the subset of overlapping users were not evenly distributed across the embedding space. We then applied a secondary alignment to the principal components to account for it. 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. We also hope to incorporate more measures to further assess temporal changes of roles.

ACKNOWLEDGMENTS

The authors would like to thank Leonardo F.R. Ribeiro for having a public implementation of *struc2vec* available and William L. Hamilton also hosting his Reddit loyalty network data publically. The work is supported by the Science Foundation Ireland under Grant Number SFI/12/RC/2289.

REFERENCES

[1] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*. 4349–4357.

[2] Hongyun Cai, Vincent W Zheng, and Kevin Chang. 2018. A comprehensive survey of graph embedding: problems, techniques and applications. *IEEE Transactions on Knowledge and Data Engineering* (2018).

[3] Sandro Cavallari, Vincent W Zheng, Hongyun Cai, Kevin Chen-Chuan Chang, and Erik Cambria. 2017. Learning community embedding with community detection and node embedding on graphs. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 377–386.

[4] Claire Donnat, Marinka Zitnik, David Hallac, and Jure Leskovec. 2018. Learning Structural Node Embeddings via Diffusion Wavelets. In *KDD*. ACM.

[5] Siobhán Grayson, Maria Mulvany, Karen Wade, Gerardine Meaney, and Derek Greene. 2017. Exploring the Role of Gender in 19th Century Fiction Through the Lens of Word Embeddings. In *International Conference on Language, Data and Knowledge*. Springer, 358–364.

[6] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 855–864.

[7] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*. 1025–1035.

[8] William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. 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)*, Vol. 1. 1489–1501.

[9] William L. Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Dan Jurafsky. 2017. Loyalty in online communities. In *International AAAI Conference on Weblogs and Social Media*, Vol. 2017. NIH Public Access, 540.

[10] Keith Henderson, Brian Gallagher, Tina Eliassi-Rad, Hanghang Tong, Sugato Basu, Leman Akoglu, Danai Koutra, Christos Faloutsos, and Lei Li. 2012. RolX: Structural Role Extraction &#38; Mining in Large Graphs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*. ACM, New York, NY, USA, 1231–1239. https://doi.org/10.1145/2339530.2339723

[11] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically Significant Detection of Linguistic Change. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World

Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 625–635. https://doi.org/10.1145/2736277.2741627

[12] Alex Leavitt. 2015. This is a Throwaway Account": Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing (CSCW ’15)*. ACM, New York, NY, USA.

[13] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[15] Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. 2016. User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest. In *ICWSM*. 279–288.

[16] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. 2016. Learning convolutional neural networks for graphs. In *International conference on machine learning*. 2014–2023.

[17] SS Nishana and Subu Surendran. 2013. Graph embedding and dimensionality reduction-a survey. *International Journal of Computer Science & Engineering Technology (IJCSET)* 4, 1 (2013), 29–34.

[18] Leonardo F.R. Ribeiro, Pedro H.P. Saverese, and Daniel R. Figueiredo. 2017. Struc2Vec: Learning Node Representations from Structural Identity, In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’17* (August 2017), 385–394. https://doi.org/10.1145/3097983.3098061

[19] Ryan A. Rossi, Brian Gallagher, Jennifer Neville, and Keith Henderson. 2013. Modeling Dynamic Behavior in Large Evolving Graphs. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM ’13)*. ACM, New York, NY, USA, 667–676. https://doi.org/10.1145/2433396.2433479

[20] Philipp Singer, Fabian Flock, Clemens Meinhart, Elias Zeitfogel, and Markus Strohmaier. 2014. Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community?. In *International World Wide Web Conference*. IW3C2, ACM, Seoul, Korea, 517–522.

[21] Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2016. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. (2016).

[22] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering* 29, 12 (2017), 2724–2743.

[23] Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. 2017. Community Preserving Network Embedding.. In *AAAI*. 203–209.

[24] Lekui Zhou, Yang Yang, Xiang Ren, Fei Wu, and Yueting Zhuang. 2018. Dynamic Network Embedding by Modeling Triadic Closure Process. In *AAAI*.

[25] Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1393–1398.