



Using Knowledge Graphs to Detect Partisanship in Online Political Discourse

Ari Decter-Frain¹ and Vlad Barash²(✉)

¹ Cornell University, Ithaca, NY 14850, USA
agd75@cornell.edu

² Graphika Technologies Inc., New York, NY 10020, USA
vlad.barash@graphika.com

Abstract. Existing methods for detecting partisanship and polarization on social media focus on either linguistic or network aspects of online communication, and tend to study a single platform. We explore the possibility of using knowledge graph embeddings to detect and analyze partisanship in online discourse. Knowledge graphs can potentially combine linguistic and network information across multiple platforms to enable more accurate discovery of a political dimension in online space. We train embeddings on heterogeneous graphs with different combinations of information text, network, single- and multi-platform information. Building on previous work, we develop a semi-supervised approach for uncovering a political dimension in the embedding space from a handful of labelled observations, and show that this method enables more accurate differentiation between liberal and conservative Twitter accounts. These results indicate that knowledge graphs can potentially be useful tools for analyzing online discourse.

Keywords: Polarization · Knowledge graphs · Graph embeddings

1 Introduction

In recent years researchers have sought to quantify online polarization via a number of different methods, using various measures of partisanship and polarization. In general, these methods focus either on the content produced by users [8, 11, 15, 20] or the networks formed by their interactions [1, 13, 19]. Fewer papers consider how focusing on each of these components may change our conclusions about polarization, or how results may generalize across platforms.

In this paper, we conduct a first exploration into the potential for knowledge graphs to enhance our understanding of political partisanship and polarization in online discourse. We fit knowledge graph embeddings (KGEs) via the library `dgl-ke` [21]. KGEs are embeddings trained on graphs with heterogeneous nodes (called entities) and edges (called relations). They encode information about network position while incorporating heterogeneous sources of information [17]. KGEs can potentially yield more detailed summaries of political discourse than

graph embeddings based on retweets, or text embeddings based on content. Furthermore, they enable the inclusion of entities from multiple social media platforms into the same network, and support analysis of how discourses play out across the social media landscape as well as how they differ between platforms.

We begin from the following question: Can unsupervised methods detect a partisanship dimension in KGEs trained on online political discourse? Furthermore, how does detection of this dimension improve as we incorporate different types of information into the graph? In addressing this question, we contribute to an ongoing line of research developing methods for quantifying controversy, partisanship, and polarization in social media discourse.

We conduct our analyses using data collected surrounding two prominent online discourses. The first pertains to disinformation about the COVID-19 pandemic (COV) and the second is about the veracity of climate change (CLM). We chose these discourses for two reasons: first, both are major topics of public discussion that generate a high volume of engagement online across multiple platforms. Chen et al. [5] found approximately 500,000 Twitter posts tagged with climate-change related hashtags in a period of just twenty days (Jan 7 2020–Jan 27 2020), while Treen et al. [18] found nearly 300,000 posts and comments. Similarly, Cinelli et al. [4] find 1,300,000 posts related to COVID-19 across multiple Gab.ai, Instagram, Reddit, Twitter, and YouTube between Jan 1 2020–Feb 14 2020, early in the pandemic. Second, COV and CLM span a wide variety of sociocultural contexts and platforms; both discourses are characterized by a high degree of political polarization [14, 18], but in neither case is the polarization confined to a specific sociocultural or political environment. We hope that our choice of discourses will allow the results of our analysis to generalize to a broad range of polarized topics.

2 Related Work

Several efforts have been made to characterize the political landscape on online social media platforms. Conover and colleagues found that a basic clustering algorithm applied to retweet and mention graphs in a political sample of Twitter data yields two large clusters [7]. Barbera developed a latent-space model for quantifying online partisanship in retweet networks [1, 2]. Interian and Ribeiro developed an empirical model of polarization that builds up from the homophily of each individual’s ego network [13]. Morales et al. [16] use a network-based approach to demonstrate how online spaces reflect the polarization in physical space. Waller and Anderson performed community detection across a full historical dataset of Reddit posts, and developed a procedure for detecting social dimensions, including partisanship in embedding space [19]. Below, we implement a variation on this procedure to measure partisanship.

Several papers focus on partisan content. Gentzkow and colleagues model polarization as the proportion of users whose partisanship can be correctly classified using the distribution of tokens from their posts [11]. Mehova et al. [15] use a combination of crowdsourcing and dictionary-based methods to characterize bias and emotion in controversial news discourse. Demsky and colleagues

combine lexical and embedding-based measures to analyze discourse surrounding mass shooting events [8]. Yan et al. [20] carry out a similar embedding exercise and validate their embedding-based measure of polarization through a stance detection task.

Some consider ways of interpreting both text and network features as indicative of polarization. Conover et al. [7] find that word frequencies differ between polarized network clusters. Chin and colleagues use both approaches to study polarization in a multiparty context [6]. Garimela and Weber follow polarization longitudinally using text- and network-based measures [10]. Garimela et al. [9] consider a method for quantifying online controversy that uses information on shared content to de-noise a retweet graph.

To our knowledge, no previous work has attempted to quantify polarization by incorporating both content and network information into a heterogeneous graph and embedding both types of information.

2.1 Data

We collected data pertaining to each of the two high profile narratives using a set of key terms to search across platforms. We used different methods for collecting data from different platforms, as follows: for Twitter, we used the public Twitter API to collect posts authored active users who posted tweets with at least one of a set of key terms, described below. Due to Twitter’s API limitations, our Twitter data collection was limited to a 30-day period. For Reddit, we used the Pushshift service [3] to collect any post with one or more of the key terms since 2011. Finally, for Telegram we constructed our own collection service to collect posts within 30 days from the collection time from public channels related to CLM or COV respectively, as identified by subject matter expert analysts who collaborated with us on this project.

For CLM, we did not discover any major public Telegram channels, so we collected data from Twitter and Reddit. The key terms for CLM are listed in the Appendix. This resulted in 799 Reddit posts between February 22, 2011 and March 15, 2022; and 155,403 Twitter posts between July 5, 2021 and August 4, 2021. For COV, we collected data from Twitter, Reddit, and Telegram. The key terms for COV are listed in the Appendix. This resulted in 1793 Reddit posts between May 25, 2011 and August 18, 2021; 3889 Telegram posts between July 15, 2021 and August 15, 2021; and 405,507 Twitter posts between May 31, 2021 and June 30, 2021.

3 Methods

3.1 Knowledge Graph Generation

We extracted knowledge graph relations from the data. Knowledge graphs consist of two basic objects: entities, which represent users who author social media posts, the context where that post appeared (e.g. a subreddit in Reddit), and the

informational content (e.g. hashtags, n-grams, URLs) present in the post text or metadata; and relations, which represent directed connections between the entities. For example, if user `@user` writes a post with hashtag `#hashtag`, we can add the entities `@user` and `#hashtag` and the relation (`@user`, `#hashtag`, “hashtag”) to the knowledge graph, where the relation is represented as an ordered triple with the “head” (the source entity of the relation), the “tail” (the target entity of the relation), and the “type” (a label for the type of the relation). Posts on different platforms sometimes, but not always, imply different relation types. For example, Twitter has a unique relation type “retweet,” but Reddit, Telegram, and Twitter all share the relation type “n-gram” (representing a user using a specific n-gram in the text of their post). We distinguish between content-relations between the author entity of a post and informational content within the same post; and network-relations between the author entity and another author entity or a context entity. We list the full set of relation types we parse from the posts described above in the Appendix. To increase graph density, we duplicated each relation to include its reverse.

Figure 1 shows a sample knowledge graph schematic using the example above. For clarity of visualization, not all the reverse relations are shown.

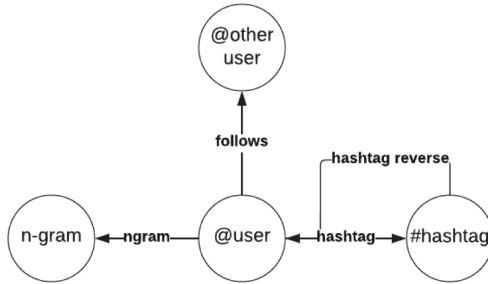


Fig. 1. Example of the relational structure of a heterogeneous knowledge graph used in this study.

As a final step, we performed K-core reduction [22] on the graphs because sparse graphs tend to inhibit convergence in dgl-ke. Beginning from $K = 10$, we computed the estimated memory requirement for training embeddings on the graph and incremented K until the estimated memory requirement fell beneath our available 32GB memory limit. We used an equation for estimating the memory burden of an embedding in a github issue [12]. Table 1 presents the breakdown of entities included in each final graph, along with the total number of relations.

Table 1 summarizes the contents of each graph, including the number of each type of entity, and the total number of relations between entities. Comparing the third and fourth column for each graph shows the effect on the total graph size from adding additional platform data. Adding data from more platforms

does add more entities to the graph, particularly accounts. However, the data we collected from other platforms generally adds few additional relations.

Table 1. Contents of each graph for which we trained embeddings. Entities are generated for each entity, and entities are linked by various relations. Min. degree describes the value used for K-core reduction to ensure each graph is as large as possible while still enabling training within memory constraints.

	Covid disinformation				Climate change Denial			
	Network	Twitter Content	All	All	Network	Twitter Content	All	All
Min. degree	10	10	20	20	10	10	10	10
N. relations	106,307	1,382,892	1,289,267	1,370,592	317,751	350,360	668,165	686,029
Account	1872	2363	2573	3538	16,926	257	16,926	17,159
Post	3279	–	1397	1397	229	–	229	229
Unigram	–	20,886	11,706	11,957	–	8,310	8309	8485
Bigram	–	8	–	–	–	–	–	2
Named entity	–	2808	1359	1395	–	1018	1018	1041
Hashtag	–	831	424	420	–	138	138	136
Media	–	225	95	95	–	3	3	3
URL	–	585	339	339	–	231	231	231

3.2 Training Embeddings

To train embeddings, we first split the arcs (triplets indicating the origin, relation type, destination of a particular relation) into training, test, and validation sets using a 90/5/5 split. We used the training and validation sets to train each embedding. We trained the embeddings using the dgl-ke library, using a fixed set of hyperparameters manually tuned to obtain adequate performance during evaluation [21]. We used the ComplEx model type because it is both flexible and efficient. We trained models for 600,000 iterations with a batch size of 1024 and a negative sample size of 512. We trained 512D embeddings using a learning rate equal to 0.5 and a regularization coefficient equal to 2.00E-6.

We evaluated the embeddings using the standard approach, by assessing their link prediction accuracy on unseen data. Link prediction asks the trained model to differentiate between a real triplet in the data (origin entity, relation, target entity), and a large number (256) of fake triplets. For each of the real and fake triplets, the model produces a probability that the triplet is real. These probabilities are ranked, and performance is measured based on where in the ranking the real triplet ends up [21]. Strong link prediction performance indicates that embeddings have encoded adequate information to identify and differentiate entities in the graph.

Table 2 presents the evaluation metrics for each trained embedding. For all models, we obtained satisfactory performance on the link prediction evaluation,

which indicates the resulting embeddings have encoded adequate information about the structure of the underlying graphs. Figure 2 summarizes the data collection, knowledge graph generation, embedding, and evaluation pipeline we describe above.

Table 2. Evaluation metrics for each trained embedding. Evaluation involves generating 512 random negative samples for each real arc, then having the model rank all 513 of these arcs in order of their probability of belonging in the graph. HITS@1 indicates the share of the test set for which the model assigns the most probability to the real arc. HITS@3 and HITS@10 are the share for which it is in the top 3 and 10, respectively. MRR is mean reciprocal rank, defined as $MRR = \frac{1}{Q} \sum_{i=1}^Q \frac{1}{rank_i}$, where Q is the number of test cases, and $rank_i$ is the rank assigned to the real triplet among the negative samples. All evaluations take place on a 5% test set unseen during training.

	COVID-19 disinformation				Climate change Denial			
	Twitter		All		Twitter		All	
	Network	Content	All	All	Network	Content	All	All
Test size	21,262	276,579	257,854	274,119	63,551	70,072	133,633	137,206
MRR	0.960	0.722	0.689	0.680	0.833	0.884	0.898	0.895
HITS@1	0.951	0.613	0.577	0.567	0.763	0.853	0.866	0.862
HITS@3	0.963	0.798	0.762	0.752	0.898	0.923	0.923	0.920
HITS@10	0.977	0.924	0.905	0.899	0.929	0.948	0.948	0.948

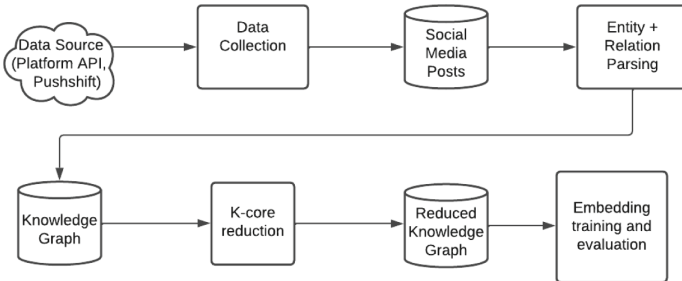


Fig. 2. Architecture diagram for data collection and knowledge graph generation.

3.3 Detecting Partisanship

To detect partisanship in knowledge graph embeddings, we apply a semi-supervised approach that builds on previous work by Waller and Anderson [19]. Our approach relies upon a small number of seed entities to construct a vector in the embedding space that represents a given social dimension. First, for each graph, we select five pairs of entities that differ in partisanship but are otherwise

similar. To select seeds, we first collected Twitter IDs for all media sources in the allsides.com media bias list¹. We joined these Twitter IDs to the entities in the graph to obtain a subset of each graph with ground-truth partisan labels. Next we computed cosine similarity between all pairs of ‘left’ and ‘right’ leaning entities and selected the five most similar pairs. We end up with five pairs per graph. Each pair contains one ‘left’ and one ‘right’ labelled entity that are otherwise relatively similar (as indicated by their proximity in the embedding space).

Next, we take the vector difference between the normalized embeddings for each pair of seeds, and average across the five pairs. The result is a 512 dimension vector that theoretically represents political partisanship in the embedding space. Once we have this vector, we can compute a partisan score for every other entity in the graph as the dot product between their normalized embeddings and the partisanship vector. Waller and Anderson show this is equivalent to taking each entity’s average similarity to the left of the partisanship vector, minus its average similarity to the right [19].

3.4 Evaluation

If our measure of partisanship accurately reflects the political partisanship of the entities in the graph, we should be able to use this measure to correctly differentiate between entities with ground-truth partisan labels. We collect a set of labelled Twitter accounts by combining two sources. First, we use the allsides.com dataset described above. Second, we collect Twitter IDs for all current members of US Congress. We recoded the labels from these two sources into three categories: Liberal, Conservative, and Center. Next, we then joined these labels into each graph. For the COV network-only and CLM content-only graphs, we obtained fewer than 50 labelled cases. Thus, we omit evaluation on these graphs. For the other six graphs, we obtained between 56 and 285 labelled cases. Using the labelled subsets of each graph, we fit logistic regressions to predict ground-truth partisanship using the partisanship scores we constructed as a predictor.

We compare performance against two baselines. First, we construct a random model where each entity’s partisanship is randomly drawn from a uniform distribution between zero and one. Second, we try using the first principal component of the embedding space as a measure of partisanship.

4 Results

Figure 3 presents two-dimensional representations of the entity embeddings obtained via principal components analysis (PCA). Each point represents an entity, with its type given by its shape, and its inferred partisanship given by its

¹ Allsides.com uses independent expert panels to assign partisan bias ratings to news outlets and journalists. <https://www.allsides.com/media-bias/ratings>.

color. For several of the graphs, some degree of polarization appears to emerge along the principal components. The principal components also appear to reflect differences between entity types. For instance, text and account entities appear to cluster somewhat separately in the graphs where the two are most heavily combined.

Interestingly, some of the text clusters appear to have strong partisan alignments. In particular, substantial clusters of blue (i.e. left-leaning) words emerge in the graphs with network and content information. These words are more ‘similar’ to the liberal seed accounts than the conservative seed accounts. This suggests that certain words, phrases, and entities tend to be used by accounts that are more similar to the liberal seeds. The knowledge graph embedding encodes this heterogeneous graphical information simultaneously.

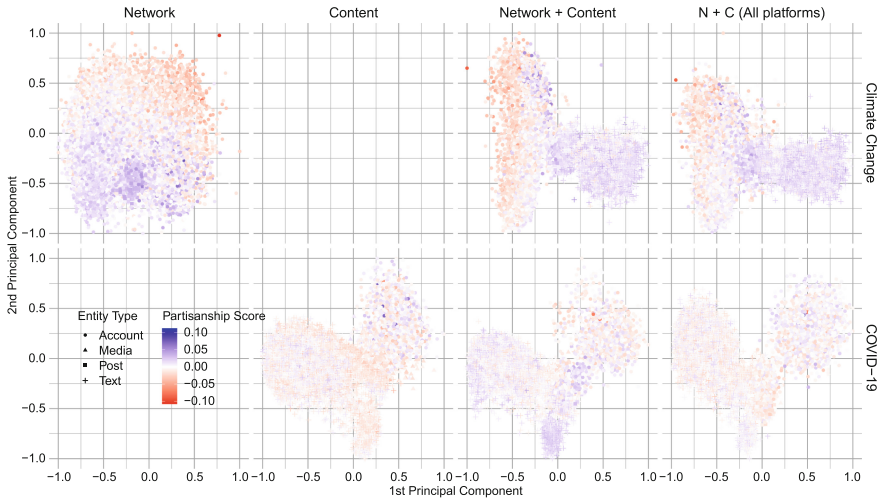


Fig. 3. Scatterplots of two-dimensional embedding space after performing knowledge graph embeddings on heterogeneous representations of online discourse. The 512 dimension embeddings are reduced to two dimensions through principal components analysis. The two omitted graphs lacked sufficient labelled data to construct/evaluate a measure of political partisanship. ‘Media’ includes URL, hashtag, and image entities. ‘Text’ includes unigrams, bigrams, and named entities.

Finally, Table 3 presents the results of our evaluation exercise. The seeded approach to measuring partisanship improves upon random and PCA baselines in every case we could measure. Improvements are largest for CLM, which may be due to the superior fit of the embeddings over these graphs. It is also notable that models with more information do not consistently outperform simpler representations of the discourse. For COV, content graphs appear better suited to detecting partisanship than more complex graphs containing content and network information. For CLM, combining network and content information yields marginally better performance than a graph with only network information.

Table 3. In-sample performance metrics for multinomial logistic regression models to trained to distinguish between liberal, conservative, and centrist media outlets using unsupervised measures of partisanship. ‘Seeded’ refers to the seed-based partisanship score. COVID-19 network and Climate Change content are omitted because these graphs lacked a sufficient number of labelled media organizations to reliably estimate performance.

		Accuracy			F1 score		
		Random	PCA	Seeded	Random	PCA	Seeded
COVID-19							
	Network	–	–	–	–	–	–
	Content	0.482	0.518	0.589	0.363	0.392	0.445
	Network + content	0.435	0.468	0.484	0.334	0.357	0.371
	N + C (All platforms)	0.484	0.419	0.484	0.370	0.321	0.370
Climate change							
	Network	0.589	0.589	0.716	0.247	0.247	0.536
	Content	–	–	–	–	–	–
	Network + content	0.589	0.600	0.740	0.247	0.289	0.563
	N + C (all platforms)	0.589	0.632	0.723	0.247	0.379	0.526

5 Discussion

In this paper we made a first attempt to apply knowledge graph embeddings to the task of learning users’ partisan affiliation via unsupervised or semi-supervised methods. We trained embeddings using graphs with different combinations of content and network information, and estimated the partisanship of every node using a seed-based procedure. These estimates were substantially better than random guessing and using the first principal component of the embeddings as a proxy for partisanship.

The data upon which we relied for our analyses have important limitations that may have factored into our results. First, several of our sampled datasets contained few accounts for which we could find ground-truth partisan labels. In two cases, this prevented us from generating and evaluating partisanship measures. Our approach of collecting data from search terms contrasts with many studies that begin from a set of labelled accounts and sample outward. This latter approach is more common and guarantees ample data for evaluation, but also risks making the problem easier by building a graph around the entities of greatest interest. Our results suggest that such a sampling strategy may lead to overly optimistic conclusions about our ability to detect partisanship.

Another challenge we ran into pertains to the magnitudes of the datasets sampled from different platforms. Although we collected data from Twitter, Telegram, and Reddit, the samples we obtained from Twitter were much larger than the other platforms. This may partly reflect the reality that Twitter contains the most discourse about the topics under study, and may also partly reflect differences in the particularities of their APIs. Here, the small sizes of the Telegram and Reddit datasets prevented us from looking at them individually, and

we could only include them as additions to the Twitter network. Future work will need to more carefully develop a framework for integrating multiplatform data into a knowledge graph to maximize information gain while accounting for platform and API idiosyncrasies.

Methods for unsupervised and semi-supervised detection of social dimensions in online discourse remain important for applied and theoretical research. This paper represents a first exploration of how knowledge graphs and knowledge graph embeddings can improve performance on this task. While our study has several important limitations, we believe it points to a promising area of research that will provide deeper, more generalizable insights into political polarization. Polarization in today’s online discourse is cross-platform as well as highly dependent on the socio-cultural context of the discourse—approaches that rely on pure text or pure network analysis are likely to draw the wrong conclusions about the mechanisms behind a polarized environment, and therefore, about the likely trajectory of its evolution.

Future research should explore solutions to the challenges we identified. This includes more systematically evaluating the effects of partisan imbalance, developing methods for incorporating multiple platforms of different sizes, and understanding when an embedding has been trained well enough to identify latent social dimensions in the data.

6 Appendix

6.1 Key Terms

For CLM, we used the following key terms: #ClimateEmergency, #EnviroScam, #ClimateAction, #climatefraud, #GreenNewDeal, #climatescare, #ActOnClimate, #NoClimateCrisis, #climatestrike, #NoClimateEmergency, #GlobalGaslighting, #climatehoax, #GlobalCooling, #ClimateReality, #climatechangehoax, #climatechangescam, #ClimateHoaxers, #ClimateScam, #GlobalWhining, #ItsCalledWeather.

For COV, we used the following key terms: #BillGatesBioTerrorist, #BillGatesVaccine, #vaxxed, #mandatoryvaccination, #forcedvaccination, #populationcontrol, #depopulation, #vaccineawareness, #BillGatesVaccine, #novaccinemandates, #medicaltyranny, #vaccineinjury, #learntherisk, #vacc, #medicalfreedom, #maskless, #uniteforfreedom, #onemillionplus, #wewillALLbethe, #vaers, #stopnewnormal, #wedonotcomply, #nomasksinclass, #novaccineforme, #destinationdepop, #nurembergtrials, #nomoremedicaltyranny, #fuckcovidjab, #vaxkseen, #nuremberg2, #ScreenB4Vaccine, #NaturalImmunity, #vaccineskill, #HumanExperiment, #UNMASKOURCHILDREN, #CrimesAgainstHumanity, #FauciLiedPeopleDied, #Agenda21.

6.2 Relation Types

See Table 4.

Table 4. Relation types by platform.

Platform	Head	Tail	Relation type
Reddit	User	Named entity	Named entity
Reddit	User	n-gram	ngram
Reddit	User	Subreddit	Subreddit
Reddit	User	URL	URL
Telegram	User	channel	Forward
Telegram	User	Named entity	Named entity
Telegram	User	n-gram	ngram
Telegram	User	User	Reply
Telegram	User	URL	URL
Twitter	User	User	Follow
Twitter	User	Hashtag	Hashtag
Twitter	User	User	Mention
Twitter	User	Named entity	Named entity
Twitter	User	n-gram	ngram
Twitter	User	User	Quote-tweet
Twitter	User	User	Reply
Twitter	User	User	Retweet
Twitter	User	URL	URL

References

1. Barberá, P.: Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Polit. Anal.* **23**(1), 76–91 (2015). <https://doi.org/10.1093/pan/mpu011>
2. Barberá, P., Jost, J.T., Nagler, J., Tucker, J.A., Bonneau, R.: Tweeting from left to right: is online political communication more than an echo chamber? *Psychol. Sci.* **26**(10), 1531–1542 (2015). <https://doi.org/10.1177/0956797615594620>
3. Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., Blackburn, J.: The Pushshift reddit dataset. In: *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)*
4. Cinelli, M., Quattrocioni, W., Galeazzi, A. et al.: The COVID-19 social media infodemic. *Sci. Rep.* **10**, 16598 (2020). <https://doi.org/10.1038/s41598-020-73510-5>
5. Chen, C., Shi, W., Yang, J., Fu, H.: Social bots’ role in climate change discussion on twitter: measuring standpoints, topics, and interaction strategies. *Adv. Clim. Change Res.* **12**(6), 913–923 (2021). <https://doi.org/10.1016/j.accre.2021.09.011>

6. Chin, A., Coimbra Vieira, C., Kim, J.: Evaluating digital polarization in multi-party systems: evidence from the German Bundestag. In: 14th ACM Web Science Conference 2022, Barcelona Spain, Jun 2022, pp. 296–301. <https://doi.org/10.1145/3501247.3531547>
7. Conover, M.D., Ratkiewicz, J., Francisco, M., Goncalves, B., Flammini, A., Menczer, F.: Political Polarization on Twitter, p. 8
8. Demszky, D., et al.: Analyzing polarization in social media: method and application to tweets on 21 Mass shootings. In: Proceedings of the 2019 Conference of the North, Minneapolis, Minnesota, 2019, pp. 2970–3005. <https://doi.org/10.18653/v1/N19-1304>
9. Garimella, K., Morales, G.D.F., Gionis, A., Mathioudakis, M.: Quantifying controversy in social media. [arXiv:1507.05224](https://arxiv.org/abs/1507.05224) [cs], Sep 2017, Accessed Apr 12, 2022 [online]. Available <http://arxiv.org/abs/1507.05224>
10. Garimella, K., Weber, I.: A Long-Term Analysis of Polarization on Twitter, p. 5
11. Gentzkow, M., Shapiro, J.M., Taddy, M.: Measuring group differences in high-dimensional choices: method and application to congressional speech. *ECTA* **87**(4), 1307–1340 (2019). <https://doi.org/10.3982/ECTA16566>
12. GitHub Issue on dgl-ke. Bus error (core dumped). <https://github.com/aws-labs/dgl-ke/issues/174>. Accessed on 07/11/2022
13. Interian, R., Ribeiro, C.C.: An empirical investigation of network polarization. *Appl. Math. Comput.* **339**, 651–662 (2018). <https://doi.org/10.1016/j.amc.2018.07.066>
14. Jiang, J., Chen, E., Lerman, K., and Ferrar, E.: Political polarization drives online conversations about COVID-19 in the United States. *Human Behav. Emerg. Technol.* **2** (2020). <https://doi.org/10.1002/hbe2.202>
15. Mejova, Y., Zhang, A.X., Diakopoulos, N., Castillo, C.: Controversy and sentiment in online news. [arXiv:1409.8152](https://arxiv.org/abs/1409.8152) [cs], Sep 2014, Accessed Apr 12, 2022 [online]. Available <http://arxiv.org/abs/1409.8152>
16. Morales, A.J., Dong, X., Bar-Yam, Y., Pentland, A.: Segregation and polarization in urban areas. *R. Soc. Open Sci.* **6**(10). <https://doi.org/10.1098/rsos.190573>
17. Nickel, M., Murphy, K., Tresp, V., Gabrilovich, E.: A review of relational machine learning for knowledge graphs. *Proc. IEEE* **104**(1), 1 (2016). <https://doi.org/10.1109/JPROC.2015.2483592>
18. Treen, K., Williams, H., O'Neill, S., Coan, T.G.: Discussion of climate change on reddit: polarized discourse or deliberative debate? *Environ. Commun.* (2022). <https://doi.org/10.1080/17524032.2022.2050776>
19. Waller, I., Anderson, A.: Quantifying social organization and political polarization in online platforms. *Nature* **600**(7888), 264–268 (2021). <https://doi.org/10.1038/s41586-021-04167-x>
20. Yan, M., Wen, X., Lin, Y.-R., Deng, L.: Quantifying content polarization on Twitter. In: 2017 IEEE 3rd International Conference on Collaboration and Internet Computing (CIC), San Jose, CA, Oct 2017, pp. 299–308. <https://doi.org/10.1109/CIC.2017.00047>
21. Zheng, D., et al.: DGL-KE: Training Knowledge Graph Embeddings at Scale. Apr 18, 2020. Accessed May 27, 2022 [online]. Available <http://arxiv.org/abs/2004.08532>
22. Seidman, S.B.: Network structure and minimum degree. *Soc. Netw.* **5**(3), 269–287 (1983). [https://doi.org/10.1016/0378-8733\(83\)90028-X](https://doi.org/10.1016/0378-8733(83)90028-X)