Unsupervised Interpretable Ideological Embedding with Variational Graph Autoencoders

ANONYMOUS AUTHOR(S)

This paper defines the novel problem of *unsupervised interpretable ideological embedding* and develops a *self-supervised algorithm*, based on variational graph auto-encoders (VGAEs), for solving this problem in polarized networks. The algorithm accepts social media users and posts as input and produces an *interpretable ideological embedding* as output. The embedding is interpretable in that it maps users and content (under broad conditions, discussed in this paper) into a joint latent space that generally satisfies three key properties, namely: (i) each axis represents a different ideological leaning, (ii) the diagonal represents neutrality, and (iii) the coordinate of an entity along an axis represents how representative an opinion it espouses is within the ideology of this particular axis. General conditions are presented and proved on both the representation learning algorithm and the input under which the produced embedding possesses the above properties. The work facilitates a number of downstream tasks, such as stance detection, stance prediction, and ideological ranking. We empirically evaluate the performance of the proposed method on multiple real-world datasets. The evaluation results show that our algorithm outperforms state-of-the-art unsupervised models in F1-score for stance-based classification, allows link prediction and user ranking within ideological groups, and extends to the scenarios of heterogeneous content.

CCS Concepts: • **Computing methodologies → Learning latent representations**; • **Information systems → Social networks**.

Additional Key Words and Phrases: Unsupervised Representation Learning, Social Networks Polarization, Interpretable Embedding, Variational Graph Auto-Encoders

ACM Reference Format:
Anonymous Author(s). 2024. Unsupervised Interpretable Ideological Embedding with Variational Graph Autoencoders. *J. ACM* 37, 4, Article 111 (September 2024), 25 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The paper defines the problem of *unsupervised interpretable ideological embedding* and offers a solution using a variational graph auto-encoder (VGAE) model, we call InfoVGAE-SL. The problem is motivated by the growing interest in the analysis of social polarization [60], where it is often desired to understand an ideological divide and rank individuals and beliefs on each side of the divide from neutral to extreme. Specifically, the question addressed in this paper is the following: given a graph of user-content interactions (such as a graph showing who posted what on social media, or who voted for what in a referendum), can one develop an *unsupervised* embedding that is *interpretable* in that the relative positions of users and items in the latent space reflect their espoused ideological leaning and (within each ideology) indicate how moderate or extreme a user or item is (relative to another) of the given ideology. Towards that end, building on a prior conference publication that introduced the notion of unsupervised belief representation learning [33], in this paper, we define the properties of an embedding that make it an *interpretable ideological embedding*. We propose and prove the broad conditions, under which the general graph embedding

0004-5411/2024/9-ART111 $15.00
https://doi.org/XXXXXXX.XXXXXXX

learning framework can produce such an interpretable embedding. We further propose a sparsity regularization term to enhance the properties of ideological separation and the proportionality between ideology and axis coordinates. We also further propose an empirical local observation compensation technique to tackle the bias of the Echo Chamber phenomenon [6]. In addition, we explore the multi-modal extensions of the proposed algorithm. An evaluation demonstrates the value of the resulting embedding in supporting downstream polarization analysis tasks.

The motivation for an *unsupervised* and *intepretable* embedding deserves further elaboration: Several standard scales are commonly used in social and political contexts to describe ideological leaning, such as the continuum from liberal to conservative, left to right, or combinations and variations thereof. For example, the Swiss political system is often depicted on a two-dimensional plain [37]. Supervised techniques have been developed that map populations into these specific latent belief spaces based on learned features such as the news outlets they consume (e.g., CNN versus Fox [23], the Guardian versus the Daily Mail [18], etc) or the words they use [24, 49]. The problem with such supervised techniques and pre-defined ideologies, however, is that they do not always transfer well across social groups, languages, and cultures. Hence, an *unsupervised* approach is sought in this paper that does not depend on domain-specific labeling.

Other approaches formulate ideology learning as entity classification or link prediction tasks [16, 22, 26, 30, 55, 65, 67], but largely neglect the interpretability of the learned representation space. In contrast, while we seek an unsupervised solution, we also want to find a latent space that is *interpretable*, so that user/item positions in the latent space are intuitively meaningful. We should mention that some statistically computable and interpretable indexes have been proposed for estimating ideological leaning [14], but they are applicable only in structured contexts where the groups are known *a priori* and their opinions are formally surveyed (e.g., via votes). In contrast, we aim to develop an unsupervised interpretable solution, whose input is social media posts.

Our solution first constructs a bipartite heterogeneous information graph of users and content items (we henceforth call *messages*), which has been shown effective to jointly model node attributes and local topology [28, 36, 45, 52, 63, 64]. Our variational graph auto-encoder (InfoVGAE-SL) then encodes both users and messages in the graph into the same disentangled latent space. We propose conditions (on the embedding loss function) under which we can show that an unsupervised embedding algorithm can produce an interpretable ideological embedding, and show that InfoVGAE-SL satisfies those conditions. A number of downstream tasks are then enabled by the properties of the proposed explainable latent space, such as (i) unsupervised *stance detection* (i.e., interpreting the stance espoused by a user or a tweet on a topic), (ii) unsupervised *stance prediction* (i.e., predicting user stance on a topic in the absence of a direct observation), (iii) unsupervised ranking, and (iv) multi-modal extensions for images and URLs.

We evaluate the performance of the proposed InfoVGAE-SL[1] on multiple real-world datasets, including Twitter datasets, a multi-modal dataset with tweets, images, and news URLs, (about the Russia/Ukraine war), and a Congress bills dataset collected from the VoteView database. The evaluation results illustrate that our method outperforms the state-of-the-art unsupervised graph embedding models, dimensionality reduction methods, and stance detection models on all datasets and produces a comparable result with state-of-the-art semi-supervised models.

The rest of this paper is organized as follows. Section 2 discusses the notion of interpretable ideological embedding, the conditions for producing such interpretability using unsupervised graph algorithms, as well as the and proofs for the theorem. Section 3 introduces the design of a concrete algorithm, InfoVGAE-SL, satisfying the conjectured feasibility conditions. Section 4 evaluates InfoVGAE-SL, demonstrating its efficacy at attaining the desired properties, as well as

[1] The code and datasets are included in the supplementary material and will be published upon acceptance.

its value to downstream inference tasks. Related work is discussed in Section 5 and the paper concludes with Section 6.

2 INTERPRETABLE IDEOLOGICAL EMBEDDING

This section defines the interpretable ideological embedding and formalizes the problem statement, one where an *unsupervised* algorithm generates a latent representation that satisfies the interpretable properties. Furthermore, we demonstrate and prove that the interpretability can be achieved by using (i) a non-negative embedding space along with (ii) decoding functions of specific properties.

2.1 Problem Formulation and Definitions

We define a social entity as $v \in \mathcal{V}$. For instance, a social entity can represent a user or post within social networks. To study the behavior of social entities in a polarized network, we introduce two key concepts: ideology and beliefs, both crucial in decision-making and information propagation among these entities. **(i) The ideology** refers to the *macroscopic* expression of a social entity's stance or preference, such as liberal or conservative ideologies in political discussions. We denote the ideologies as $\{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_d\}$, where d is the number of ideologies. d is a pre-defined hyper-parameter for specific applications, which can be selected based on heuristics. **(ii) The beliefs** $b \in B$ represent the *microcosmic* expression of a social entity's ideology, encompassing detailed views on various topics. For example, an entity with a liberal ideology might hold a belief b_1 supporting the rights of immigration and another belief b_2 advocating for environmental protection. Consequently, each ideology is characterized by a finite set of beliefs (i.e., a *belief system*), $\mathcal{I}_k = \{b_1, b_2, \ldots\}$. Similarly, each social entity also maintains its own belief system $\boldsymbol{b}_v = \{b_1, b_2, \ldots\}$.

Definition 1 (Ideology and Belief System): *The ideology $\mathcal{I}$, and the individual-level belief system $\boldsymbol{b}_v$ of a social entity v are defined as finite sets of beliefs:*

$$\mathcal{I} = \{b_1, b_2, \ldots\}, \ \boldsymbol{b}_v = \{b_1, b_2, \ldots\}. \tag{1}$$

A belief b is considered more *representative* and *extreme* within an ideology $\mathcal{I}$ if it is (i) more *popular* within that ideology and (ii) more *discriminative* across ideologies (i.e., less popular in other ideologies). For instance, in highly polarized scenarios, user interaction across ideologies is minimal, leading to more extreme beliefs. Conversely, in a less polarized scenario, most beliefs are widely accepted and equally popular among all ideologies and social entities, resulting in less extreme (neutral) beliefs. Based on this, we can further define a function $R^{\mathcal{I}}(v)$ to quantify the extreme of a social entity v within an ideology $\mathcal{I}$, according to its beliefs $\boldsymbol{b}_v$.

Definition 2 (Ideology Group): *The ideology group $\mathcal{F}_k$ (also known as ideology faction) is defined as a set comprising social entities who are the most representative in the k-th ideology $\mathcal{I}_k$.*

$$\mathcal{F}_k = \{v \in \mathcal{V} \mid R^{\mathcal{I}_k}(v) \geq R^{\mathcal{I}_i}(v), \forall i \in \{1, 2, \ldots, d\}\} \tag{2}$$

In real-world settings, social entities are exposed to a mix of topics and interact with each other, leading to information propagation and the polarization phenomenon. Research in social psychology [12] suggests that social entities with similar belief systems $\boldsymbol{b}_v$ tend to reach consensus. In social networks, this consensus often translates into positive *social interactions* such as *"likes"*, *"follows"*, and *"re-posts"*. Based on these observations, this paper proposes a general definition of an interpretable ideology embedding space that can automatically discover and categorize the ideologies $\mathcal{I}$ as well as the ideologies of social entities $\boldsymbol{b}_v$ based on the *social interactions*. Building on these concepts, we define the interpretable ideology embedding as follows:

Definition 3 (Interpretable Ideological Embedding): *The Interpretable Ideological Embedding is an encoding that maps social entities into a embedding space that fulfills the following three conditions:*

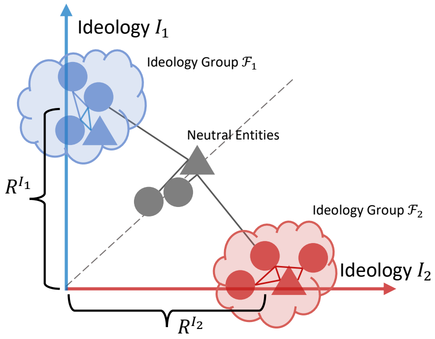


Fig. 1. Demonstrative example of a 2D interpretable ideological embedding. The circles and triangles represent social entities v such as users and messages. Each axis represents an ideology $\mathcal{I}$ with its corresponding ideology group $\mathcal{F}$. The coordinates represent the extreme measure $R^{\mathcal{I}}$ of each entity.

(1) *The embedding space is a rectangular coordinate space where each axis represents a distinct ideology $\mathcal{I}$ and its corresponding ideology group, $\mathcal{F}$.*
(2) *The coordinate of a social entity along an axis indicates the extent of its extreme with the ideology associated with that axis. Specifically, the coordinates for social entities along the axis of an ideology $\mathcal{I}$ are proportional to $R^{\mathcal{I}}()$.*
(3) *The diagonal of the axis represents neutrality, meaning that we have $R^{\mathcal{I}_i}(b_{\bar{v}}) \approx R^{\mathcal{I}_j}(b_{\bar{v}})$, $\forall i, j$ for any neutral entity $\bar{v}$.*

An interpretable ideological embedding is *unsupervised* if it is derived through an algorithm that operates without access to pre-annotated ideology labels, does not interpret the beliefs b, and is unaware of the functions $R^{\mathcal{I}}()$. An example of the ideological embedding is shown in Fig. 1.

2.2 Conditions for Interpretability

In this section, we examine the general conditions necessary to ensure interpretability of the latent space within the context of unsupervised graph embedding models like variational graph auto-encoders (VGAE) [29]. Consider a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ represents entities and their attributes, and $\mathcal{E}$ denotes the edges. Unsupervised graph embedding models strive to learn a latent representation $\boldsymbol{z} \in \mathbb{R}^d$ for each entity by maximizing the likelihood of link existence, where d is the dimension of the latent space. We denote $\boldsymbol{Z} \in \mathbb{R}^{|\mathcal{V}| \times d}$ as the embedding matrix, with $\boldsymbol{z} \in \boldsymbol{Z}$. The primary optimization objective of these models is expressed in Equation (3), which maximizes a similarity between connected entity pairs (when $e_{i,j} = 1$) while minimizing it between unconnected pairs (when $e_{i,j} = 0$), utilizing a pairwise decoding function $\xi : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow [0, 1]$.

$$\max_{\boldsymbol{\theta}} \ \Pi_{i,j} \ \xi(z_i, z_j)^{e_{i,j}} \left[1 - \xi(z_i, z_j)\right]^{1-e_{i,j}}, \tag{3}$$

where $\boldsymbol{\theta}$ represents the model parameters. In VGAE models, which include a trainable encoder, $\boldsymbol{\theta}$ encompasses the parameters of GCN encoders (i.e., $\boldsymbol{z}$ is parameterized by $\boldsymbol{\theta}$). In models that lack encoders, $\boldsymbol{\theta} = \boldsymbol{Z}$ acts as the trainable latent embedding. Based on this graph learning framework, we propose specific conditions to ensure the interpretability of the generated embedding space.

Theorem 1 (Interpretability Conditions): *An unsupervised graph embedding learning model produces an interpretable ideological embedding space if the following conditions are satisfied:*

(1) *The embedding space is non-negative, i.e. $\boldsymbol{Z} \geq 0$.*
(2) *The decoding function $\xi(z_i, z_j) \in \mathbb{R}_{[0,1]}$ is symmetric, smooth, continuous, and differentiable.*

(3) $\xi(z_i, z_j) = 0$ *if* z_i *is orthogonal to* z_j *(i.e., when* $z_i z_j^\top = 0$*); here,* z_i *and* z_j *can be* $\mathbf{0}$.
(4) $\xi(z_i, z_j) \approx 1$ *when* z_i *is close to* z_j *under a distance criterion.*
(5) $\xi(z_i, z_j)$ *is positively correlated with* $|z_i|$ *and* $|z_j|$*, i.e.* $\frac{\partial \xi}{\partial |z_i|}, \frac{\partial \xi}{\partial |z_j|} \geq 0$.

Various choices are possible for the decoding function ξ as well as the loss function used to optimize Equation (3). For instance, a suitable decoding function ξ satisfying these conditions can be a *scaled sigmoid inner product* $\xi(z_i, z_j) = 2\sigma(z_i z_j^\top) - 1$, and a widely used loss function is *binary cross entropy*. We demonstrate that the three properties of the generated embedding Z, as outlined in Definition 3, are satisfied when these conditions are met.

2.3 Proof of Interpretability under the Defined Conditions

Consider binary cross entropy loss as an illustrative example. Assume there are d ideology groups, denoted as $\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_d$. With ξ as the decoding function, the loss function for learning unsupervised graph embedding is expressed as

$$\mathcal{L} = -\frac{1}{|\mathcal{V}|^2} \sum_{i,j} \Big[e_{i,j} \log \xi(z_i, z_j) + (1 - e_{i,j}) \log \big(1 - \xi(z_i, z_j)\big) \Big]. \tag{4}$$

For both the *positive* edges ($e_{i,j} = 1$) and *negative* (unconnected) edges ($e_{i,j} = 0$), the summations can be categorized into two parts: *intra-ideology group* or *inter-ideology group.*

For brevity, we use $\mathbf{1}_{i,j}^{\mathcal{F}}$ to denote:

- $\mathbf{1}_{i,j}^{\mathcal{F}} = 1$ if i and j belong to the same ideology group, i.e., $\exists k \in \{1, 2, \ldots, d\}$, s.t. $i, j \in \mathcal{F}_k$.
- $\mathbf{1}_{i,j}^{\mathcal{F}} = 0$ if i and j belong to two different ideology groups, i.e., $i \in \mathcal{F}_k, j \in \mathcal{F}_r, k \neq r$.

We further decompose Equation (4) as follows,

$$\begin{aligned}\mathcal{L} = -\frac{1}{|\mathcal{V}|^2} \sum_{i,j} \Big[& \underbrace{\mathbf{1}_{i,j}^{\mathcal{F}} e_{i,j} \log \xi(z_i, z_j)}_{\text{(i) intra-group, positive}} + \underbrace{\mathbf{1}_{i,j}^{\mathcal{F}} (1 - e_{i,j}) \log \big(1 - \xi(z_i, z_j)\big)}_{\text{(ii) intra-group, negative}} \\ & + \underbrace{(1 - \mathbf{1}_{i,j}^{\mathcal{F}}) e_{i,j} \log \xi(z_i, z_j)}_{\text{(iii) inter-group, positive}} + \underbrace{(1 - \mathbf{1}_{i,j}^{\mathcal{F}})(1 - e_{i,j}) \log \big(1 - \xi(z_i, z_j)\big)}_{\text{(iv) inter-group, negative}} \Big].\end{aligned} \tag{5}$$

Assuming the interpretability conditions outlined in Theorem 1 are satisfied, we now prove the three properties defined in Definition 3 as follows:

Proof of Definition 3(1) Consider p is a concrete social entity associated with a polarized ideology group $\mathcal{F}_p$ (i.e., $p \in \mathcal{F}_p$). Taking the partial derivative of the loss function in Equation (5) with respect to $\xi(z_i, z_j)$ and letting $i = p$, we have,

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \xi(z_p, z_j)} = -\frac{1}{|\mathcal{V}|^2} \sum_{j} \Big[& \underbrace{\frac{\mathbf{1}_{p,j}^{\mathcal{F}} e_{p,j}}{\xi(z_p, z_j)}}_{\text{(i) intra-group, positive}} - \underbrace{\frac{\mathbf{1}_{p,j}^{\mathcal{F}} (1 - e_{p,j})}{1 - \xi(z_p, z_j)}}_{\text{(ii) intra-group, negative}} \\ & + \underbrace{\frac{(1 - \mathbf{1}_{p,j}^{\mathcal{F}}) e_{p,j}}{\xi(z_p, z_j)}}_{\text{(iii) inter-group, positive}} - \underbrace{\frac{(1 - \mathbf{1}_{p,j}^{\mathcal{F}})(1 - e_{p,j})}{1 - \xi(z_p, z_j)}}_{\text{(iv) inter-group, negative}} \Big].\end{aligned} \tag{6}$$

In a polarized scenario, the number of positive connections *within* each ideology group significantly exceeds the negative connections, that is, $|\{(i, j) \mid \mathbf{1}_{i,j}^{\mathcal{F}} e_{i,j} = 1\}| \gg |\{(i, j) \mid \mathbf{1}_{i,j}^{\mathcal{F}} (1 - e_{i,j}) = 1\}|$ for the entity pairs $(i, j) \in \mathcal{V} \times \mathcal{V}$. Meanwhile, positive connections *across* different ideology groups are far fewer than negative connections, i.e., $|\{(i,j) \mid (1-\mathbf{1}^{\mathcal{F}}_{p,j})(1-e_{p,j})=1\}| \gg |\{(i,j) \mid (1-\mathbf{1}^{\mathcal{F}}_{p,j})e_{p,j}=1\}|$. Therefore, we can conclude that the gradient in Equation (6) is dominated by terms (i) and (iv) in the polarized scenarios. When the gradient descent converges for (i) and (iv), we have,

$$\frac{\mathbf{1}^{\mathcal{F}}_{p,j} e_{p,j}}{\xi(z_p, z_j)} \longrightarrow \min\,, \quad \frac{(1-\mathbf{1}^{\mathcal{F}}_{p,j})(1-e_{p,j})}{1-\xi(z_p,z_j)} \longrightarrow \min\,. \tag{7}$$

This approaching suggests

- $\xi(z_p, z_j) \approx 1$ for any entity j of the same ideology group, i.e. $\forall j \in \mathcal{F}_p$
- $\xi(z_p, z_j) \approx 0$ for any entity j of a different ideology group, i.e. $\forall j \in \mathcal{F}_r, r \neq p$

Referencing Theorem 1(3), and given the conclusion above where $\xi(z_p, z_j) \approx 0$, $\forall j \in \mathcal{F}_k, k \neq p$, it follows that z_p is **near-orthogonal** to z_j when j is from a different ideology group. Considering $Z \geq 0$ is a non-negative rectangular coordinate system as introduced in Theorem 1(1), each ideology group will be aligned along a distinct axis upon the convergence of optimization. □

Proof of Definition 3(2) Similarly, consider two entities p and p^+ within the same ideology group, i.e., $p, p^+ \in \mathcal{F}_p$, the embeddings of which are z_p and z_{p^+}, respectively. The partial derivative of loss function in Equation (5) with respect to $|z_p|$ (or to $|z_{p^+}|$ by replacing p with p^+) is given by:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial |z_p|} = -\frac{1}{|\mathcal{V}|^2}\sum_j \Big[& \underbrace{\frac{\mathbf{1}^{\mathcal{F}}_{p,j} e_{p,j}}{\xi(z_p,z_j)} \frac{\partial \xi(z_p,z_j)}{\partial |z_p|}}_{\text{(i) intra-group, positive}} - \underbrace{\frac{\mathbf{1}^{\mathcal{F}}_{p,j}(1-e_{p,j})}{1-\xi(z_p,z_j)} \frac{\partial \xi(z_p,z_j)}{\partial |z_p|}}_{\text{(ii) intra-group, negative}} \\ & + \underbrace{\frac{(1-\mathbf{1}^{\mathcal{F}}_{p,j}) e_{p,j}}{\xi(z_p,z_j)} \frac{\partial \xi(z_p,z_j)}{\partial |z_p|}}_{\text{(iii) inter-group, positive}} - \underbrace{\frac{(1-\mathbf{1}^{\mathcal{F}}_{p,j})(1-e_{p,j})}{1-\xi(z_p,z_j)} \frac{\partial \xi(z_p,z_j)}{\partial |z_p|}}_{\text{(iv) inter-group, negative}} \Big].\end{aligned} \tag{8}$$

As previously introduced, in the polarized scenarios the (i) and (iv) in Equation (8) contain significantly more terms than (ii) and (iii). In this context, (ii) and (iii) act as the regularization for (i) and (iv), respectively. Assume p^+ is more extreme within $\mathcal{F}_p$ than p, meaning that p^+ is more likely to agree with same-ideology entities and oppose cross-ideology entities. Thus, $|\{j \in \mathcal{V} \mid \mathbf{1}^{\mathcal{F}}_{p^+,j} e_{p^+,j} = 1\}|$ and $|\{j \in \mathcal{V} \mid (1-\mathbf{1}^{\mathcal{F}}_{p^+,j})(1-e_{p^+,j}) = 1\}|$ are larger for p^+ than for p. Consequently, the regularization (ii) and (iii) for p^+ (i.e. for $\frac{\partial \mathcal{L}}{\partial |z_{p^+}|}$) contain even fewer terms than those for p (i.e. for $\frac{\partial \mathcal{L}}{\partial |z_p|}$). Referencing the condition specified in Theorem 1(5), we have $\frac{\partial \xi(z_p,z_j)}{\partial |z_p|} \geq 0$ and $\frac{\partial \xi(z_{p^+},z_j)}{\partial |z_{p^+}|} \geq 0$. Therefore, we illustrate that the regularization for p^+ is weaker than that for p.

During the gradient descent, the update of $|z_p|$ follows the direction of $|z_p| := |z_p| - \eta \frac{\partial \mathcal{L}}{\partial |z_p|}$ (or respectively $|z_{p^+}| := |z_{p^+}| - \eta \frac{\partial \mathcal{L}}{\partial |z_{p^+}|}$ for $|z_{p^+}|$), where η is a non-negative step size. Moreover, given the near-orthogonal property as proven above, we have $\xi(z_p, z_j) \approx 0$ for inter-ideology entities, i.e., for the terms (iv). Therefore, when the optimization is almost converged (i.e., when the embeddings are aligned with axis), the more effective terms are (i) and (ii), which are non-negative. With a smaller regularization in (ii) for $|z_{p^+}|$, we can conclude that $|z_{p^+}| > |z_p|$. Since z_{p^+} and z_p are aligned along the same ideology axis of $\mathcal{F}_p$, their coordinates along $\mathcal{F}_p$ are equivalent to $|z_{p^+}|$ and $|z_p|$, respectively. Therefore, we demonstrate that a larger coordinate on the ideology axis represents an entity with more extreme beliefs. □

Proof of Definition 3(3) In the polarized scenario, the near-orthogonal property proven previously indicates that the embedding z is sparse, with nearly only one element being non-zero when optimization converges. Consider n as a neutral entity with its embedding denoted as z_n.

The behavior of this neutral entity suggests equal interaction with entities from different ideology groups. The loss function for z_n can be re-written as the following equation:

$$\mathcal{L}|_{i=n} = -\sum_{k=1}^{d}\sum_{j\in\mathcal{F}_k}\Big[\underbrace{e_{n,j}\log\xi(z_n,z_j)}_{\text{(i) positive}} + \underbrace{(1-e_{n,j})\log\big(1-\xi(z_n,z_j)\big)}_{\text{(ii) negative}}\Big]. \tag{9}$$

Given the behavior of the neutral entity, the count of edges $|\{(n,j)\mid e_{n,j}=1, j\in\mathcal{F}_k\}|$ to entities j of $\mathcal{F}_k$ is approximately equal across all ideology groups, i.e., for $\forall k\in\{1,2,\ldots,d\}$. We denote $|\{(n,j)\mid e_{n,j}=1, j\in\mathcal{F}_k\}| = C_1$ and correspondingly $|\{e_{n,j}\mid e_{n,j}=0, j\in\mathcal{F}_k\}| = C_2$, where C_1 and C_2 are constants. For brevity, we assume a complete polarization for all the other entities except for z_n, in which we have: (1) The sizes of different ideology groups are similar, i.e. $|\mathcal{F}_k|\approx|\mathcal{F}_r|, \forall k,r$. (2) There are no inter-ideology interaction.

In this case, the coordinate value of any z_j $(j\neq n)$ along its corresponding ideology axis $\mathcal{I}_k$ (i.e. the axis of $\mathcal{F}_k$) is nearly equal. meaning that $z_j\approx\alpha\boldsymbol{e}_k, \forall j\in\mathcal{F}_k, \forall k$, where $\boldsymbol{e}_k$ is the unit vector of the axis of $\mathcal{F}_k$. Let z_n^m denote the m-th element of z_n, we derive:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial z_n^m} &= -\sum_{k=1}^{d}\sum_{j\in\mathcal{F}_k}\Big[\frac{e_{n,j}}{\xi(z_n,z_j)}\frac{\partial\xi(z_n,z_j)}{\partial z_n^m} - \frac{1-e_{n,j}}{1-\xi(z_n,z_j)}\frac{\partial\xi(z_n,z_j)}{\partial z_n^m}\Big]\\ &\approx -\sum_{k=1}^{d}\Big[\frac{C_1}{\xi(z_n,\alpha\boldsymbol{e}_k)}\frac{\partial\xi(z_n,\alpha\boldsymbol{e}_k)}{\partial z_n^m} - \frac{C_2}{1-\xi(z_n,\alpha\boldsymbol{e}_k)}\frac{\partial\xi(z_n,\alpha\boldsymbol{e}_k)}{\partial z_n^m}\Big],\end{aligned} \tag{10}$$

Setting $\frac{\partial\mathcal{L}}{\partial z_n^m}=0$ for every m in $\mathbb{N}_{[1,d]}$ and considering symmetry, the solution to these d equations will result in $z_n^1 = z_n^2 = \ldots = z_n^d$, indicating that the diagonal of the axis represents neutrality with respect to the ideology of social entities. □

We have demonstrated that the three properties of interpretability in Definition 3 are achieved when the specified conditions are met and the optimization fully converges to the global minimum. Meanwhile however, other local minima may also exist during the optimization of the embedding space. Consequently, there remains a *challenge* in model design to avoid getting trapped in undesirable local minima. For instance, as will be introduced in Section 3.4, imposing regularization that favors sparse representations can guide the model toward a solution where more entities align with individual ideological axes, potentially avoiding other minima. With this understanding, we now introduce an empirical model in the following Section 3, which is demonstrated to successfully achieve unsupervised interpretable ideological embedding with multiple enhancement modules.

3 UNSUPERVISED INTERPRETABLE IDEOLOGICAL REPRESENTATION LEARNING WITH INFOVGAE-SL

Expanding the problem statement presented in Section 2, we model users and messages by a *Bipartite Heterogeneous Information Network (BHIN)* [57] given by a graph, $\mathcal{G}=\{\mathcal{V},\mathcal{E}\}$, where the number of vertices is $|\mathcal{V}|=N$ and the number of edges is $|\mathcal{E}|=M$. The number of vertex types is 2. The number of edge types is R. The BHIN can also be written as $\mathcal{G}=\{\{\mathcal{V}_1,\mathcal{V}_2\},\{\mathcal{E}_1,\mathcal{E}_1,...,\mathcal{E}_R\}\}$. Each possible edge from the i^{th} vertex to the j^{th} vertex is denoted as $e_{ij}\in\mathcal{E}$ with a weight value w_{ij}. The dimensions of the adjacency matrix, A, are $(|\mathcal{V}_1|+|\mathcal{V}_2|)\times(|\mathcal{V}_1|+|\mathcal{V}_2|)$, where $A_{i,j}=w_{ij}$. We model $\mathcal{G}$ as *undirected*, where $\langle v_i,v_j\rangle\equiv\langle v_j,v_i\rangle$. Heterogeneity of edges allows expressing multiple types of actions. For instance, in a voting example, different edge types may represent the actions of voting *Yea*, *Nay*, or *Abstain*. The problem becomes converting an input BHIN into the (maximum likelihood) latent representation $\boldsymbol{z}$ for users and messages with explainable properties.

In this section, we describe the **Info**rmation-Theoretic **V**ariational **G**raph **A**uto-**E**ncoder with **S**parsity Regularization and **L**ocal Observation Compensation (**InfoVGAE-SL**) model which maps the users and messages into the defined interpretable ideological latent space for polarization analysis. The overall structure of InfoVGAE-SL is shown in Fig. 2. Importantly, the training of the proposed InfoVGAE-SL model minimizes Equation (4) with $\xi(z_i, z_j) = 2\sigma(z_i z_j^\top) - 1$, where σ is the sigmoid function. Observe that the above substitution satisfies the four desired properties of ξ (needed for unsupervised interpretable ideological embedding). Below, we describe the detailed design and applications of the proposed InfoVGAE-SL model.

3.1 Non-Negative Inference Model (Encoder)

The inference model takes a constructed BHIN as the input, denoted as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$. We use $\boldsymbol{A}$ to denote the adjacency matrix of $\mathcal{G}$ with self loops. $\boldsymbol{X} \in \mathbb{R}^{N\times F}$ is the input feature.

We use the Graph Convolutional Network [28] (GCN) of L layers as the network architecture of the encoder. Assume in the l^{th} layer, the hidden state of GCN is $\boldsymbol{G}^{(l)} = GCN^{(l)}(\boldsymbol{A}, \boldsymbol{X}), \boldsymbol{G}^{(l)} \in \mathbb{R}^{N\times d_l}$, where d_l is the dimension of hidden state in the l^{th} layer. $\boldsymbol{G}^{(1)} = \boldsymbol{X}$ is the input feature matrix. $\boldsymbol{X}$ could be initialized as identity matrix $\boldsymbol{I}$ if there is no available feature. The GCN layer is formulated as

$$\boldsymbol{G}^{(l)} = \gamma\left(\widetilde{\boldsymbol{A}}\boldsymbol{G}^{(l-1)}\boldsymbol{W}^{(l-1)}\right),\ (2 \leq l \leq L-1) \tag{11}$$

where $\widetilde{\boldsymbol{A}} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$ is the symmetrically normalized adjacency matrix. $\boldsymbol{D}$ is diagonal degree matrix with the diagonal element $\boldsymbol{D}_{k,k} = \sum_{i=1}^{N} \boldsymbol{A}_{k_i}$. $\boldsymbol{W}_l \in \mathbb{R}^{d_l \times d_{l+1}}$ is the weight matrix in the l^{th} layer. γ denotes the activation function. We use Rectified Linear Unit (ReLU) as the activation function in our model. For the BHIN with multiple edge types ($R > 1$), we extend GCN with Relational GCN [44].

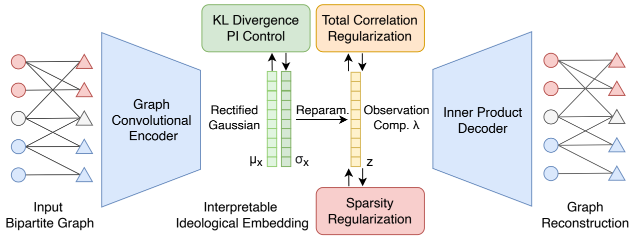


Fig. 2. InfoVGAE-SL consists of non-negative graph convolutional encoder, inner product decoder, and latent space control modules. It encodes the user-message bipartite graph into a explainable latent space and then reconstructs it. KL divergence control, total correlation regularization, sparsity regularization, and observation compensation techniques are proposed to enhance the disentanglement and interpretability of latent space.

Assume $\boldsymbol{Z} \in \mathbb{R}^{N\times T}$ is the latent matrix. T is the dimension of target representation space. $\boldsymbol{z}_i$ is the latent vector of the i^{th} node. The inference model is defined as:

$$q(Z|\boldsymbol{A}, \boldsymbol{X}) = \prod_{i=1}^{N} q(z_i|\boldsymbol{A}, \boldsymbol{X}),\ q(z_i|\boldsymbol{A}, \boldsymbol{X}) \sim \mathcal{N}_{+}(z_i|\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2), \tag{12}$$

where posterior is assumed to follow the rectified Gaussian Distribution $\mathcal{N}_{+} = max(\mathcal{N}, 0)$ [51]. $\boldsymbol{\mu} = \widetilde{\boldsymbol{A}}\boldsymbol{G}^{(L-1)}\boldsymbol{W}_{\boldsymbol{\mu}}^{(L-1)}$ is the matrix of mean vectors $\boldsymbol{\mu}_i$. $\log\boldsymbol{\sigma} = \widetilde{\boldsymbol{A}}\boldsymbol{G}^{(L-1)}\boldsymbol{W}_{\boldsymbol{\sigma}}^{(L-1)}$ is the matrix of standard deviation vectors $\boldsymbol{\sigma}_i$. They share the same hidden state $\boldsymbol{G}^{L-1}$.

The use of a rectified Gaussian Distribution ($\mathcal{N}_{+}$) to form the target latent space is not accidental. While the positions espoused by opposing ideologies are generally in conflict, they are not necessarily *opposite* and, for that matter, not always mutually exclusive. For example, some might favor funding the military, while others might favor funding environmental research. Some might believe in doing both. This is in contrast to *stance detection*, where stance on some topic is either positive, neutral, or negative. Thus, in representing *systems of belief*, we remove the negative side of each latent axis, thereby forcing the disentanglement of different ideologies onto *different axes* (as opposed to a single axis with positive versus negative values). This representation offers better compositionality. For example, individuals who mix and match elements of different belief systems (e.g., strongly believe in funding both the military and the environment) can now be better represented. Polarization is more prominent when these individuals and contents align more closely with the axes in latent space (and away from the origin) as opposed to being closer to the diagonals. Section 4.7 shows that our non-negative target space indeed helps separate polarized nodes into corresponding axes.

3.2 Generative Model (Decoder)

We use a linear inner product decoder. Its generative model can be formulated as:

$$p(\boldsymbol{A}|\boldsymbol{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{N} p(\boldsymbol{A}_{i,j}|\boldsymbol{z}_i, \boldsymbol{z}_j),\ p(\boldsymbol{A}_{i,j}|\boldsymbol{z}_i, \boldsymbol{z}_j) = 2\sigma(\boldsymbol{z}_i^T \boldsymbol{z}_j) - 1, \tag{13}$$

where $\sigma(\cdot)$ is the logistic sigmoid function. This inner product decoder enhances the explainability of our latent space. In the geometric space, $\boldsymbol{z}_i^T \boldsymbol{z}_j$ is defined as the multiplication between the norm of the projection of $\boldsymbol{z}_i$ over $\boldsymbol{z}_j$ and the norm of $\boldsymbol{z}_j$. While we maximize $p(\boldsymbol{A}_{i,j}|\boldsymbol{z}_i, \boldsymbol{z}_j)$, we are forcing the latent vectors $\boldsymbol{z}_i, \boldsymbol{z}_j$ to be parallel if $\boldsymbol{A}_{i,j} = 1$ and orthogonal otherwise. As a result, different systems of belief will be associated with different orthogonal axes.

3.3 Total Correlation Regularization

Inspired by Total Correlation (TC) [66] in information theory, we design a total correlation regularization module in InfoVGAE-SL to encourage the model to disentangle the latent variables. The total correlation is a generalization of mutual information for multiple random variables. By penalizing the total correlation term, the model is able to learn a series of statistically independent latent variables, thereby making the belief representations interpretable.

Let z_i^k denote the k^{th} dimension of the latent vector for the i^{th} node, the total correlation regularizer is defined as

$$\mathcal{L}_{TC}(\boldsymbol{z}_i) = D_{KL}\left[q(\boldsymbol{z}_i)\|\prod_{k=1}^{d} q(\boldsymbol{z}_i^k))\right], \tag{14}$$

which can be interpreted as the divergence between joint distribution and independent distribution. However, this KL divergence is intractable, since the entire dataset is required to evaluate every $q(\boldsymbol{z}_i)$ with direct Monte Carlo estimate. There are two main approaches in the literature to tackle this problem: (i) Decompose the KL divergence and then use Monte Carlo sampling to compute the independent probability [13, 20, 71]; (ii) Train a discriminator to approximate the KL divergence with density-ratio trick [27, 41, 54]. In this paper, we train a discriminator $\Phi(\boldsymbol{z}_i)$ to discriminate whether a sample $\boldsymbol{z}_i$ is from the joint distribution $q(\boldsymbol{z}_i)$ or independent distribution $\prod_{k=1}^{d} q(\boldsymbol{z}_i^k)$. The total correlation thus can be approximated by

$$\mathcal{L}_{TC}(\boldsymbol{z}_i) \approx \mathbb{E}_{\boldsymbol{z}_i \sim q(\boldsymbol{z}_i)}[\log(\Phi(\boldsymbol{z}_i)) - \log(1 - \Phi(\boldsymbol{z}_i))]. \tag{15}$$

To jointly train the discriminator and VGAE model, in each training step of VAE, we sample a batch of $\boldsymbol{z}_i$ from $q(\boldsymbol{z}_i)$ as positive samples (joint distribution), and generate the negative samples (independent distribution) with the independence trick [4]. For every latent dimension, we randomly permute the values across different items in the batch. Then the parameters of discriminator can be optimized via the maximum likelihood estimation.

3.4 Sparsity Regularization

As introduced in Section 2.1, we expect the learned representation Z to have a more clear separation and ranking for nodes of different ideologies. Since our latent space is non-negative, jointly enhancing the sparsity can also enhance the *orthogonality* of Z. L1 regularization has been proven to be effective in encouraging sparsity [70]. In this paper, we propose to apply L1 regularization on the representations Z to improve the ideological separation as well as the ideology extreme proportionality of learned representations. The regularization term is formulated as,

$$\mathcal{L}_{SP} = \gamma \sum_{i=1}^{N} \sum_{j=1}^{T} |\mathbf{Z}_{i,j}|, \tag{16}$$

Where γ is the weight for sparsity regularization. Note that $|\mathbf{Z}_{i,j}| = \mathbf{Z}_{i,j}$ when $\mathbf{Z}_{i,j} \geq 0$. This regularization term penalizes the value of $\mathbf{Z}_{i,j}$ and therefore eliminates small values of $\mathbf{Z}_{i,j}$ to be 0.

This sparsity regularization term is also effective in improving the explainability of the axis coordinates introduced in Section 2.1. For a pair of connected nodes ($A_{i,j} = 1$) which fall in the same ideological axis $\mathcal{I}^{\mathcal{F}}$, we assume their belief values (non-zero coordinates) are b_1 and b_2. In this case, we have $z_i^T z_j \approx b_1 b_2$ in Equation (13). Applying the sparsity regularization term is equivalent to setting up a constraint ξ for $b_1 + b_2$ so that $b_1 + b_2 \leq \xi$. According to AM–GM inequality, we have $b_1 b_2 \leq (b1 + b_2)^2/4 \leq \xi^2/4$ where the first equation holds when $b_1 = b_2$ and the second equation also holds when b_1 and b_2 reach their maximums. Therefore, for the positive edges, while maximizing the inner product $z_i^T z_j$, the regularization will enforce b_1 to be close to b_2. In this way, the nodes of similar beliefs will be placed nearby, which enhances the statements in Section 2.

3.5 Empirical Local Observation Compensation

The Echo Chamber or Information Cocoon phenomenon arises when platforms selectively expose users to content aligned with their expressed interests [19, 69], such as when recommendation systems promote content based on users' interests and ideologies. As a result, users are limited to a *local observation* of content that aligns with their intrinsic beliefs. Sociological research has also shown that Echo Chamber effects are becoming increasingly prevalent on online social networks, especially in political discourse [6, 48]. This local observation bias can significantly affect existing ideological prediction models on social networks, particularly those relying on social interactions. In this section, we propose a Local Observation Compensation (LOC) parameter, λ, to empirically mitigate the effects of the Echo Chamber phenomenon, thereby generating more unbiased ideological representations and improving performance on downstream tasks.

Building on Equation (4), we introduce a weighting term $w = \lambda \frac{|\mathcal{N}|}{|\mathcal{P}|}$ to balance the sparsity of the graph by adjusting the ratio between the positive samples $\mathcal{P} = \{(i, j) \mid e_{i,j} = 1\}$ and negative samples $\mathcal{N} = \{(i, j) \mid e_{i,j} = 0\}$:

$$\mathcal{L} = -\frac{1}{|\mathcal{V}|^2} \sum_{i,j} \Big[w \cdot e_{i,j} \log \xi(z_i, z_j) + (1 - e_{i,j}) \log \big(1 - \xi(z_i, z_j)\big) \Big]. \tag{17}$$

In previous works [29], w is typically set as the ratio $w = \frac{|\mathcal{N}|}{|\mathcal{P}|}$ to balance positive and negative samples, addressing the sparsity issue. In this case, $w > 1$ when the interaction graph is sparse. However, this method assumes that users have a *global observation* of all content. This assumption may hold for certain interaction graphs, such as the Voteview dataset, where each congress voter reviews all bills (messages) and makes decisions. In such cases, voters have global visibility of all content. However, in datasets like Twitter, users only have a local observation of a subset of content, as no user can view all recent tweets. User observations depend on their activities and the recommendation system's selections. Here, the negative samples $\mathcal{N}$ are ambiguous, as they may indicate either disagreement with the content or a lack of exposure to the corresponding messages.

To address this local observation issue, we introduce a compensating hyper-parameter $\lambda > 1$ to further emphasize the importance of $\mathcal{P}$, which is less noisy than $\mathcal{N}$. Thus, the weighting term is redefined as $w = \lambda \frac{|\mathcal{N}|}{|\mathcal{P}|}$. Empirically, the optimal value of λ can be determined using a validation subset of the target dataset. In Section 4.6, we explore the impact of λ on various datasets and the resulting improvements in downstream tasks.

3.6 Joint Training with KL Divergence PI Control

The design of VAEs often suffers from KL-vanishing, also called posterior collapse, in that the value of KL-divergence becomes zero during model training. This implies over-fitting and entails a failure to generalize from training data. This phenomenon is especially significant for graph representation learning models [56]. We introduce the Proportional Integral (PI) control module to tackle the KL vanishing problem and ensure a disentangled latent space. The PI controller [46] can dynamically tunes $\beta(t)$ to manipulate the KL-divergence based on the difference between the actual value and the target value during model training. The control process can be formulated as:

$$\beta(t) = \frac{K_p}{1 + \exp(e(t))} - K_i \sum_{j=0}^{t} e(j), \tag{18}$$

where $e(t)$ is the difference between the target KL_{set} and the actual KL-divergence at training step, t. K_p and K_i are positive hyper-parameters of the designed PI controller. The error pushes $\beta(t)$ in a direction that causes KL-divergence to approach the target. When the KL-divergence is too small, $e(t)$ becomes positive, causing the output, $\beta(t)$, of the PI controller to decrease, thereby boosting the actual KL-divergence to higher values. This mechanism encourages InfoVGAE to learn a more informative latent representation.

The overall objective of InfoVGAE-SL includes optimizing the evidence lower bound (ELBO) of VAE while simultaneously minimizing the total correlation and applying sparsity regularization, which can be formulated as

$$\begin{aligned} &\mathbb{E}_{Z\sim q(Z|\boldsymbol{A},\boldsymbol{X})}\left[\log p(\boldsymbol{A}|\boldsymbol{Z})\right] - \beta(t) D_{KL}\left[q(\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{X}) \| p(\boldsymbol{Z})\right] \\ &\quad -\lambda \mathbb{E}_{Z\sim q(Z)}\left[\log(\Phi(\boldsymbol{Z})) - \log(1-\Phi(\boldsymbol{Z}))\right] - \mathcal{L}_{SP}, \end{aligned} \tag{19}$$

where the first two terms are the ELBO objective with PI control variable $\beta(t)$. The last term is the total correlation regularizer introduced in Section 3.3. The joint training process of VAE, total correlation regularizer, and PI control module brings additional benefits for InfoVGAE-SL to learn a disentangled and informative latent representation.

3.7 Downstream Tasks

Once a latent representation is learned, several downstream tasks become possible. The easiest is to determine the stance or ideological polarity espoused by users and messages (depending on whether the input data comprises opinions on one topic, as in stance, or views on a number of different topics, thus revealing a system of beliefs). The disentangled latent space produced by the InfoVGAE-SL offers a simpler way to separate such stances or polarities. Fig. 6 shows an example of learned representations in Voteview dataset. Every axis is associated with a different latent belief system. To select the dominant ones, we choose the axes with the largest accumulated projection values over all the data points. We then use point coordinates along those axes as measures of alignment with the corresponding ideologies. Thus, we can classify the polarity of a user or message simply by considering the axis where it has the largest coordinate (without using a clustering algorithm). We can also predict the likelihood that a user agrees with an message from their proximity in the latent space. We can also rank users and messages on each axis by the corresponding coordinate values to determine how strongly they espouse the corresponding ideology. Examples of these applications are presented in Section 4.

4 EXPERIMENTS

In this section, we evaluate the performance of the proposed InfoVGAE-SL on a wide spectrum of downstream applications built upon our ideological embedding. The experiments are conducted based on Python 3.6.2 and Pytorch 1.7.0 framework, on a device with 128-core CPU, 256GB RAM, and 3 NVIDIA 3090 GPUs.

4.1 Datasets

The following data sets are used in the evaluation.

US Election 2020: We collected a real-world dataset via the Twitter search API using keywords *{president, election, trump, biden}*. A total of 1, 149, 438 tweets were collected about the US presidential election from *Dec 4, 2020* to *Dec 23, 2020*. The dataset captures debate about the legitimacy of the election process and includes many opinion statements about Donald Trump, Joe Biden, and related events in their campaign. Individual tweet cascades (i.e., a tweet and its retweets) were called messages, one per cascade. We asked human graders to manually label 844 most popular messages for evaluation as either pro-Trump or pro-Biden. Among our labeled messages, there were 237 messages supporting Trump and 607 messages supporting Biden.

Eurovision: This public dataset is about the annual Eurovision Song Contest [2]. It was used for polarity detection. The background is that Susana Jamaladinova (Jamala) from Ukraine won the Eurovision 2016 contest with a song named *1944*. This song ignited controversy due to political connotations and possible violations of Eurovision rules. Some users opposed the song quoting Eurovision rules that prevent politically motivated entries. Others applauded Jamala for her rendition of the plight of an ethnic minority, who suffered (presumably by Russian hands) as described in the song. In this dataset, 600 messages were manually labeled pro-Jamala and 239 were labeled anti-Jamala.

Russia-Ukraine War (Multi-Modal): The dataset is collected from Twitter that includes users' attitudes and reactions on Russia-Ukraine war from *May 1, 2022* to *Aug 8, 2022*. A total of 961352 tweets where collected with 14661 URLs and 3603 images. As before, 11280 the most prolific users and 5020 the most popular messages were retained. In this dataset, 155 messages support Ukraine and 409 messages supporting Russia are manually labeled for evaluation.

Voteview: We collected the voting data of the 105^{th} Congress (that held office towards the end of the 90s) from the Voteview [31] database that documents U.S. Congress voting records. Our collected data contains information on 444 congressmen from different parties, 1166 bills, and the voting records. Since most congressmen are Republican or Democrat, we only consider them for polarization analysis. For ground truth, we label the congressmen with their party affiliations, and label the bills with the majority party affiliation of congressmen who voted *Yea*.

4.2 Baselines

We compare the proposed InfoVGAE-SL with 9 baselines:

Non-Negative Matrix Factorization (NMF) [2]: This is an unsupervised approach that uncovers polarization based on factorizing the matrix of users and their messages. Unlike the VGAE, matrix factorization breaks down an observations matrix into an encoder (matrix) and a decoder (matrix) that are *both linear*.

Belief Structured Matrix Factorization (BSMF) [68]: This is an enhancement to Non-Negative Matrix Factorization handling situations where different community belief systems partially overlap.

Graph Convolutional Networks (GCN) [28]: The regular GCN is a semi-supervised model that encodes node features and graph structure into representations with a small set of labels. For a fair comparison, we adopt an unsupervised GCN with a softmax classifier after GCN layers for link prediction during training.

Stance Detection (Stance) [17]: An unsupervised stance detection model which uses texts, hashtags, and mentions to build features, and maps users into a low dimensional space with UMAP.

DeepWalk [42]: This method learns a latent social representation by modeling a series of short random walks. It maps the nodes into a relatively small number of dimensions, capturing neighborhood similarity and community membership.

Signed Polarity Propagation (sPP) [1]: This method represents opinions of individuals with signed bipartite networks and formulates polarity analysis as a classification task. A linear algorithm is proposed to learn the polarity labels exploiting network effects.

TIMME (TIMME-Sup) [67]: TIMME is a semi-supervised multi-task and multi-relational embedding model. TIMME first models the social networks as a heterogeneous graph and jointly trains several link prediction tasks and an entity classification task to extract a latent representation of users. We use TIMME-Sup to refer to the TIMME-hierarchical architecture in the original implementation.

Unsupervised TIMME (TIMME-Unsup) [67]: An unsupervised variant of TIMME-hierarchical without the supervision of entity classification task.

InfoVGAE [33]: The previous conference version of the proposed InfoVGAE-SL model without sparsity regularization and local observation compensation modules.

4.3 Qualitative Examples

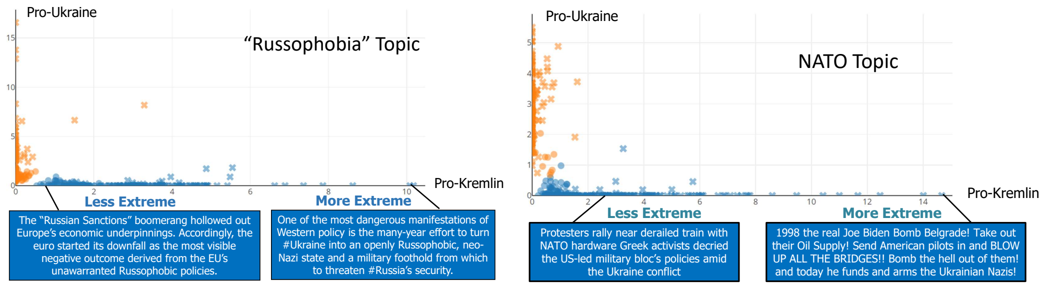
In this section, we present qualitative examples of the nature of the embedding produced by our algorithm. Specifically, for purposes of qualitative illustration, we filter the Russia-Ukraine data set by topic, and present two examples of (debated) subtopics claimed by pro-Kremlin messaging: (i) the allegations of “Russophobia" – the Kremlin complaint that Russian nationals are not treated fairly abroad, and (ii) an exchange of accusations over the role of NATO in promoting conflict. Each debated subtopic has two sides: the pro-Kremin views and the pro-Ukraine views. The unsupervised algorithm maps individuals (denoted by circles) and messages (denoted by crosses) into a two-dimensional space, where each axis represents a different side of the debate. As discussed in Section 2, we expect the mapping to satisfy the three conditions of unsupervised interpretable ideological embedding. Axis labels are then manually added. In a later section (on *polarity detection*), we quantitatively evaluate the accuracy of ideological separation in the latent space. A qualitative illustration of the embedding for the two sides is shown in Fig. 3 for the two debates, respectively. The color-coding denotes ideological affiliation. To verify whether messages further from the origin are indeed more “extreme", we show two message examples in each figure, drawn from the pro-Kremlin axis. (Actual text is presented unfiltered; the reader should exercise discretion as some text may be offensive to some readers.) While the degree of severity is subjective, the examples show that messages further from the origin seem to carry more ‘incrimination’ of the other side, compared to those closer to the origin (and as such may be perceived as more offensive). In a later section, we evaluate the algorithm’s ideological ranking ability more rigorously based on a data set of US Congress members whose ground-truth ideology is known.

Table 1. Evaluation of clustering results for polarity and stance detection on three Twitter datasets and the Voteview dataset. Note that *Stance Detection is Twitter-specific and not applicable for Voteview. †sPP is developed for signed political bipartite graph and only applicable for the Voteview dataset. ‡TIMME-Sup is trained in a supervised manner. Stance Detection and TIMME models only support the prediction of users.

Dataset: US Election 2020							
Model Name	User Prec.	User Recall	User F1	Msg. Prec.	Msg. Recall	Msg. F1	Purity
NMF[2]	0.4275	0.8235	0.5628	0.4130	0.6786	0.5135	0.6313
BSMF[68]	0.6970	0.6866	0.6917	0.3818	0.7778	0.5122	0.6959
GCN[28]	0.5699	0.7910	0.6625	0.3455	0.7037	0.4634	0.6512
DeepWalk[42]	0.9310	0.8060	0.8640	**0.8824**	0.5556	0.6818	0.8571
Stance[17]	**0.9429**	0.6226	0.7500	-	-	-	0.8240
TIMME-Unsup[67]	0.9322	0.8209	0.8730	-	-	-	0.7822
InfoVGAE[33]	0.9333	0.8358	0.8819	0.6667	0.8148	0.7333	0.8599
InfoVGAE-SL (Ours)	0.9344	**0.8507**	**0.8906**	0.7419	**0.8519**	**0.7931**	**0.8794**
TIMME-Sup[67]‡	1.0000	0.8333	0.9091	-	-	-	-
Dataset: Eurovision 2016							
Model Name	User Prec.	User Recall	User F1	Msg. Prec.	Msg. Recall	Msg. F1	Purity
NMF	0.3202	0.9286	0.4762	0.3142	0.5352	0.3960	0.7123
BSMF	0.5337	0.6786	0.5975	0.2866	0.5352	0.3733	0.7248
GCN	0.3113	**0.9429**	0.4681	0.2918	0.8594	0.4356	0.7135
DeepWalk	0.3028	**0.9429**	0.4583	0.2895	**0.8867**	0.4365	0.7217
Stance	0.4280	0.9134	0.5829	-	-	-	0.6947
TIMME-Unsup	0.9513	0.7778	0.8556	-	-	-	0.7865
InfoVGAE	0.9649	0.7857	0.8661	0.8447	0.5312	0.6523	0.8842
InfoVGAE-SL (Ours)	**0.9905**	0.7820	**0.8739**	**0.8954**	0.5805	**0.7044**	**0.8900**
TIMME-Sup‡	0.9907	0.7852	0.8760	-	-	-	-
Dataset: Russia Ukraine War 2022							
Model Name	User Prec.	User Recall	User F1	Msg. Prec.	Msg. Recall	Msg. F1	Purity
NMF	0.9510	0.9814	0.9659	0.6019	0.7806	0.6797	0.8976
BSMF	0.5431	0.7397	0.6264	0.2988	0.6903	0.4171	0.6899
GCN	0.6546	0.9907	0.7883	0.3909	0.9483	0.5536	0.7709
DeepWalk	0.7180	0.8463	0.7769	0.6706	0.8057	0.7320	0.7329
Stance	**0.9654**	0.9834	0.9743	-	-	-	0.9037
TIMME-Unsup	0.8860	0.9834	0.9322	-	-	-	0.8373
InfoVGAE	0.9611	0.9891	0.9749	0.8928	**0.9677**	0.9287	0.9713
InfoVGAE-SL (Ours)	0.9608	**0.9911**	**0.9757**	**0.9090**	**0.9677**	**0.9374**	**0.9737**
TIMME-Sup‡	0.9679	0.9955	0.9815	-	-	-	-
Dataset: Voteview*							
Model Name	User Prec.	User Recall	User F1	Msg. Prec.	Msg. Recall	Msg. F1	Purity
NMF	0.9952	0.9763	0.9856	0.4957	0.9971	0.6622	0.8451
BSMF	0.9718	0.9764	0.9741	0.4826	0.9971	0.6504	0.8383
GCN	0.4742	0.9528	0.6332	0.4203	0.8563	0.5639	0.6149
DeepWalk	0.9952	0.9763	0.9856	0.4922	0.9971	0.6591	0.8451
sPP[1]†	0.9718	0.9764	0.9741	0.8427	0.8621	0.8523	0.9430
TIMME-Unsup	0.9765	0.9432	0.9595	-	-	-	0.8827
InfoVGAE	0.9952	**0.9811**	0.9881	**0.9878**	0.9339	**0.9601**	0.9828
InfoVGAE-SL (Ours)	**1.0000**	**0.9811**	**0.9905**	0.9817	**1.0000**	0.9581	**0.9835**
TIMME-Sup‡	0.9850	0.9924	0.9887	-	-	-	-

4.4 Polarity and Stance Detection

After learning the latent representations, the simplest downstream application is stance and/or polarity detection. While with InfoVGAE-SL classification can be done based on coordinate values, for the compared baselines we apply a K-Means clustering algorithm to predict the polarity of



The "Russian Sanctions" boomerang hollowed out Europe's economic underpinnings. Accordingly, the euro started its downfall as the most visible negative outcome derived from the EU's unawarranted Russophobic policies.

One of the most dangerous manifestations of Western policy is the many-year effort to turn #Ukraine into an openly Russophobic, neo-Nazi state and a military foothold from which to threaten #Russia's security.

Protesters rally near derailed train with NATO hardware Greek activists decried the US-led military bloc's policies amid the Ukraine conflict

1998 the real Joe Biden Bomb Belgrade! Take out their Oil Supply! Send American pilots in and BLOW UP ALL THE BRIDGES!! Bomb the hell out of them! and today he funds and arms the Ukrainian Nazis!

Fig. 3. Ideological Embedding of users and messages (marked as circles and crosses) in the Russia/Ukraine conflict. The discussion topic of the left is about "Russophobia", and the right is about the role of NATO.

users and messages. The number of clusters is set to 2, since in all the datasets, there are two dominant sides to the conflict. The weight γ for sparsity regularization is set as 5.0×10^{-6}, 1.0×10^{-6}, 5×10^{-7}, and 5×10^{-7} for Election, Eurovision, Russia-Ukraine War, and VoteView datasets. In this experiment, we conduct clustering separately for users and messages for a fair comparison since some baselines only produce user representations. We use common information retrieval metrics including precision, recall, and F1-score to evaluate the performance. We also calculate the purity metric introduced in [17, 39]. To compute purity, each cluster is assigned to the class which is the most frequent in the cluster. The purity is measured by counting the number of correctly assigned data points and divided by the total number of points N. Assume k is the number of clusters, $\Omega = \{\omega_1, \omega_2, \cdots, \omega_k\}$ is the set of predicted clusters and $\mathbf{C} = \{c_1, c_2, \cdots, c_k\}$ is the set of classes, the purity score is defined as $purity = \frac{1}{N} \sum_{i=1}^{k} \max_j |\omega_i \cap c_j|$, where c_j is the classification which has the maximum count for cluster ω_i. A higher purity represents a cleaner separation.

The evaluation results are shown in Table 1. The results for the Twitter datasets illustrate stance detection (pro-Jamala versus against, pro-Trump versus against, pro-Russia versus against). The results for Voteview dataset illustrate polarity separation (Democrats versus Republicans). We observe that InfoVGAE-SL achieves the highest user F1-Score and purity scores on all the four datasets, except that on Voteview dataset, InfoVGAE achieves the highest message F1 score. The proposed InfoVGAE and InfoVGAE-SL model can jointly map users and messages into a disentangled latent space that mutually enhances each other. In addition, the KL divergence control, total correlation, and sparsity regularization helps make the learned representations to be sparse and disentangled.

DeepWalk has a higher F1 score and purity than most other baselines on the US Election and Russia-Ukraine War dataset but does not work well on Eurovision, because the Eurovision dataset contains more noise, causing the Deepwalk to generate many redundant clusters. Stance and TIMME-Unsup are state-of-the-art stance detection and social graph representation learning methods. Unlike InfoVGAE-SL, they can only produce the representations of users due to their specific model frameworks. Therefore, we only compare the metrics of users. The F1 score of TIMME-Unsup is comparative but slightly lower than the InfoVGAE-SL model on all datasets. The Stance Detection model produces comparative result at Russian Ukraine War dataset, but a relatively low F1 score on the other datasets. The competitiveness of them is attributed to their additional utilization of mention or hashtag data. Some baselines produce a higher recall or precision but a smaller F1 score, compared with InfoVGAE-SL. The reason is the clustering algorithm is confused by the embedding and mistakenly cluster most points together as one category. In the Voteview dataset, the adjacency matrix represents the voting record of congressmen. It is more dense and contains less noise. Therefore, most models produce a high User F1-score over 0.97. However, the baselines still cannot achieve a comparable Bill F1-score as InfoVGAE or InfoVGAE-SL, because their bill representations are less informative and harder to be separated. InfoVGAE-SL is specialized for improving the interpretability of ideological embedding with the proposed sparsity regularization and Local Observation Compensation (LOC) techniques, it outperforms InfoVGAE in most F1 and purity scores for Polarity and Stance Detection.

We also compare the result of InfoVGAE-SL with a supervised model. InfoVGAE-SL produces a very comparative result with TIMME-Sup, while InfoVGAE-SL is an unsupervised method. In US Election, Eurovision datasets, and Russia-Ukraine War datasets, the gap of user F1 score is also narrowed into 1.85%, 0.21%, and 0.58%, respectively. In the Voteview dataset, our InfoVGAE-SL model even outperforms TIMME-Sup by 0.18% in user F1 score. It's reasonable that supervised methods outperform unsupervised ones on most datasets. However, the evaluation result of InfoVGAE-SL is reaching the upper bound of all unsupervised methods, with a special design and control for latent space distribution. The proposed model also outperforms TIMME-Unsup in the PureP dataset introduced in [67] with 0.70% and 0.73% higher accuracy and F1 score, as introduced in the conference paper [33].

4.5 Case Study of Stance Separation

Many existing stance detection models only support the stance evaluation of users, such as Stance Detection [17] and TIMME [67], whereas the design of InfoVGAE-SL enables us to further separate the stances of messages made by users. We show the top 5 messages separated by stance by our unsupervised algorithm (to get a feel for the data at hand). As mentioned above, each axis in the disentangled latent space produced by InfoVGAE-SL is associated with a different ideology. Based on this observation, we sort and rank the messages by their coordinates on each axis and report (for illustration) the top-5 messages (with the largest coordinate) on each axis. Results are shown in Table 2 (for the Jamala data set). Column labels are added manually.

Table 2. Tweets with the Top-5 highest polarities in Eurovision dataset, ranked by the coordinate value of corresponding axis in the latent space of learned belief representations by InfoVGAE-SL.

Pro-Jamala	Anti-Jamala
RT @jamala: Thank you for your love! #jamala #eurovision #jamala1944 #eurovision_ukraine #cometogether	Jamala's 1944: Song for Nazi.
Incredible performance by #Jamala, giving Crimean Tatars, suffering persecution & abuse, reason to celebrate.	I must say I feel a little sorry for @jamala, from the start simply a tool in the West's #CrimeanTatars campaign.
President awarded @jamala title of the People's Artist of Ukraine.	if Jamala's singing of "Our Crimea", a totally non-political song, wasn't against the rules - why is every video of it removed now?
#CantStopTheFeeling #Eurovision Congrats @jamala #Ukraine!!	Ukraine's Eurovision winner, Jamala, is so angry with Russia that she appeared at Sochi's New Year party ($$)
This scene will give me goosebumps until the day I die. Thanks for such a masterpiece @jamala.	@jtimberlake @Eurovision @jamala The day Eurovision REALLY went political. What a shame #Eurovision.

Now how, in Table 2, *Pro-Jamala* tweets are expressing gratitude and congratulate Jamala. The *Anti-Jamala* tweets criticize Jamala's song as political and a violation of the rules of the Eurovision contest.

4.6 Ablation Study of Local Observation Compensation

In section 3.5 we propose the Local Observation Compensation (LOC) parameter λ to tackle the bias introduced by the Echo Chamber effect. In this section, we study and empirically estimate the optimal λ for the polarity and stance detection downstream task on different datasets. The F1 score of users under different λ is shown in Fig. 4. In the Voteview dataset, since all voters have

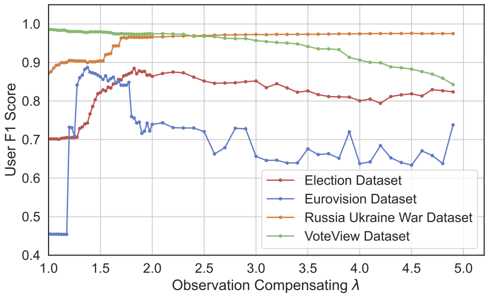


Fig. 4. The validation F1 score of users under different Local Observation Compensation (LOC) parameter λ. On the Twitter datasets, the best F1 score is achieved when $\lambda > 1$ (LOC enabled). On VoteView dataset (green curve), since there is no local observation problem, the best F1 score is achieved when $\lambda = 1$ (LOC disabled).

the global observation of congress bills, applying no compensating ($\lambda = 1$) will lead to the best evaluation results. In the Twitter datasets of US Election and Eurovision, the model achieves the highest F1 score under the optimal $\hat{\lambda} = 1.825$ and $\hat{\lambda} = 1.375$. On these two datasets, when $1 < \lambda < \hat{\lambda}$ the observation bias of $\mathcal{N}$ leads to the under-convergence of the model. When $\lambda > \hat{\lambda}$, the model over-emphasizes the positive edges $\mathcal{P}$ and therefore produces a biased result. On Russia Ukraine Dataset, the local observation problem is the most significant, therefore a relatively larger λ always brings improvement to the model. In the experiment sections, we use grid-search to find the best dataset-specific λ for the InfoVGAE-SL model.

4.7 Ideology Analysis

As defined in Section 2.1, the interpretable ideological embedding should be able to separate the different ideological groups, as well as represent the ideological strength with coordinates. In this section, we evaluate the models' ability to generate a proportional coordinates with the ground-truth ideological strength of users.

Table 3. Ideology evaluation in three Twitter datasets and VoteView dataset. We report the Kendall correlation scores and the cosine similarity scores. A higher Kendall correlation and cosine similarity represents a better match between ground-truth ideology value and the predicted polarity.

Dataset: US Election 2020			Dataset: Eurovision 2016		
Model Name	Kendall Score	Cosine Similarity	Model Name	Kendall Score	Cosine Similarity
NMF	0.3703	0.3948	NMF	0.6147	0.5122
BSMF	0.4772	0.4064	BSMF	0.3616	0.4464
GCN	0.4714	0.4143	GCN	0.3919	0.4362
DeepWalk	0.8072	0.6387	DeepWalk	0.5098	0.3924
TIMME-Unsup	0.7366	0.3841	TIMME-Unsup	0.6378	0.6908
InfoVGAE	0.7038	0.6585	InfoVGAE	0.7786	0.7740
InfoVGAE-SL	**0.8358**	**0.6852**	**InfoVGAE-SL**	**0.7944**	**0.7815**
Dataset: Ukraine Russian War			Dataset: Voteview		
Model Name	Kendall Score	Cosine Similarity	Model Name	Kendall Score	Cosine Similarity
NMF	0.9316	0.9231	NMF	0.8035	0.9647
BSMF	0.4318	0.3480	BSMF	0.8041	0.9650
GCN	0.8707	0.5904	GCN	0.7227	0.4004
DeepWalk	0.9098	0.9035	DeepWalk	0.8301	0.9655
TIMME-Unsup	0.7613	0.8376	TIMME-Unsup	0.8150	0.9478
InfoVGAE	0.9363	0.9256	InfoVGAE	0.8345	**0.9688**
InfoVGAE-SL	**0.9414**	**0.9659**	**InfoVGAE-SL**	**0.8812**	0.9663

In Voteview database, an ground-truth ideology value [10] generated by DW-NOMINATE [38] algorithm is provided, which represents the static ideological position of each Congress member across the course of their career. This ideology value is calculated based on large amounts of data in history and can be used as the benchmark for congressman's ideological leanings. In this scoring system, the ideology value is positive for Republican congressmen and negative for Democratic congressmen. A larger absolute value means a more deeply entrenched position. In Twitter datasets, we apply a similar way to measure the ground-truth ideology value of one user by calculating the proportion of tweets' ideologies of the given users. In the binary case, the ideology value is calculated by $(|t^+| - |t^-|)/(|t^+| + |t^-|)$, where $|t^+|$ and $|t^-|$ are the number of tweets of positive and negative ideologies (manually annotated) for a given user. The ideology value is within $[-1, 1]$.

Ideally, the produced polarity value by InfoVGAE-SL (i.e., the projection of a point on its dominant ideology axis) should be strongly correlated with the ground-truth ideology value. To quantitatively evaluate the correctness of the polarity ranking produced by our InfoVGAE-SL model, we use Kendall Rank Correlation Coefficient to evaluate the correlation between the ranking sequences of polarity and ideology. We also use Cosine Similarity to evaluate the similarity between polarity and ideology values. Note that for baselines, the polarity prediction of ideology $\mathcal{I}_k$ is computed by $\max_u\{\|z_u - c_k\|_2\} - \|z_i - c_k\|_2$ for the i-th user, where c_k is the KMeans clustering centroid of $\mathcal{I}_k$. For InfoVGAE-SL we can directly use the coordinates as the polarity prediction without any clustering algorithm. The result is shown in Table 3. InfoVGAE-SL achieved the highest Kendall correlation on all datasets and the highest cosine similarity on Election, Eurovision, and Russia-Ukraine War datasets. The gain of Kendall correlation is attributed to the explainable representation and especially the ranking enhancement with the sparsity regularization module.

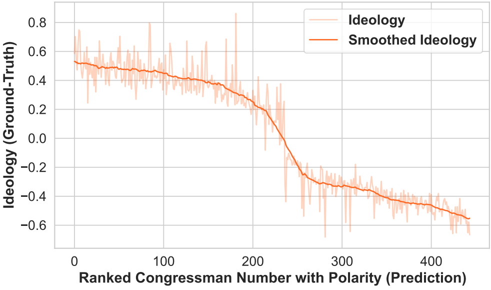


Fig. 5. Ideology of ranked Congressmen. The Smoothed ideology is that smoothed with a slide window. The overall trend of the ideology is monotonic, which means the ranking predicted by InfoVGAE-SL (x-axis) is consistent with the ground-truth ideology (y-axis).

On VoteView dataset, we further visualize the ideology of ranked congressmen in Fig. 5. The horizontal axis is the order of congressmen ranked with their polarity value. In the figure, clear correlations are seen with the ground-truth ideology value. In addition, we show the top 5 retrieved congressmen of the highest latent value of polarity computed by InfoVGAE-SL in Table 4. These individuals espouse the most extreme positions. We look up and report their ground-truth ideology scores in table, showing that they are indeed extreme compared to the party average ideology score in the last row. This table offers further intuition into the ranking quality with computed polarity.

4.8 Stance Prediction

We do not actually present detailed results on stance *prediction* here, but rather present evidence that it should be possible to predict stance from the embedding. Fig. 6 is a 2D projection of the

Table 4. Top-5 Congressmen with the highest polarity retrieved by InfoVGAE-SL. Ideology represents ground-truth of static ideological position.

Democratic	Ideology	Republican	Ideology
McDermott, Jim	−0.666	Sessions, Pete	0.586
Oliver, John Walter	−0.577	Stump, Robert Lee	0.703
Filner, Bob	−0.652	Ryun, Jim	0.547
Owens, Major	−0.569	Paxon, L. William	0.472
Lewis, John R.	−0.589	Armey, Richard Keith	0.635
Democratic Average	−0.376	*Republican Average*	0.402

InfoVGAE-SL's latent representations of congressmen and bills. It shows ground truth on passed and failed bills as well as ground truth on the party responsible for passing or failing them. It also shows the ground truth party affiliation of Congress members. The diagonal separates the two belief systems. Note how most bills above the diagonal (in the Republican space) are either passed by Republicans or failed by Democrats. Similarly, most bills below the diagonal (in the Democrat area) are either passed by Democrats or failed by Republicans. The figure shows that the latent representation of bills learned by InfoVGAE-SL indeed predicts the parties which will vote for/against them.

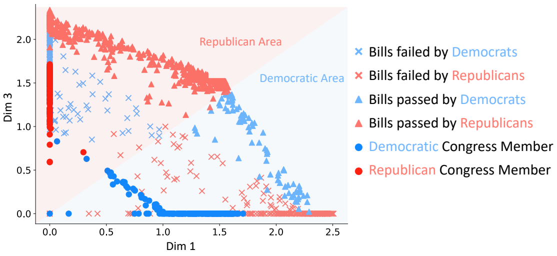


Fig. 6. Learned ideological embedding in Voteview dataset. We projected the 3D representations into 2D to present and explain the results of polarity detection by InfoVGAE-SL. The colors and shapes of data points represent ground-truth.

Table 5. Results of ablation studies. The top five rows of each dataset show the results after disabling the total correlation module, PI controller, rectified Gaussian, sparsity regularization, and joint learning.

Dataset: US Election 2020				Dataset: Eurovision 2016			
Model Name	User F1	Tweet F1	Purity	Model Name	User F1	Tweet F1	Purity
Without Total Corr.	0.8780	0.7273	0.8660	Without Total Corr.	0.8625	0.5100	0.8196
Without Rect. Gaussian	0.8594	0.4938	0.7753	Without Rect. Gaussian	0.6209	0.5093	0.7144
Without KL Control	0.8358	0.6545	0.8097	Without KL Control	0.8506	0.5045	0.8175
Without Sparsity Reg.	0.8819	0.7333	0.8599	Without Sparsity Reg.	0.8661	0.6523	0.8842
Separate-Learning	0.8615	0.7111	0.8597	Separate-Learning	0.8497	0.6449	0.8249
InfoVGAE-SL	**0.8906**	**0.7931**	**0.8794**	**InfoVGAE-SL**	**0.8739**	**0.7044**	**0.8900**
Dataset: Ukraine Russian War				Dataset: Voteview			
Model Name	User F1	Tweet F1	Purity	Model Name	User F1	Tweet F1	Purity
Without Total Corr.	0.9597	0.8783	0.9545	Without Total Corr.	0.9881	0.9518	0.9807
Without Rect. Gaussian	0.9487	0.7831	0.9184	Without Rect. Gaussian	0.9881	0.9309	0.9751
Without KL Control	0.9585	0.8783	0.9541	Without KL Control	0.9858	0.6578	0.8440
Without Sparsity Reg.	0.9749	0.9287	0.9713	Without Sparsity Reg.	0.9881	**0.9601**	0.9828
Separate-Learning	0.9630	0.7975	0.9287	Separate-Learning	0.9791	0.9218	0.9659
InfoVGAE-SL	**0.9757**	**0.9374**	**0.9737**	**InfoVGAE-SL**	**0.9905**	0.9581	**0.9835**

4.9 Ablation Studies

We further conduct ablation studies to explore the impact of proposed modules on polarity detection. We keep all other experimental settings unchanged except for the ablation module. The experimental results are shown in Table 5.

Effect of Total Correlation Module: We remove the discriminator for total correlation regularization and remove the total correlation term in our objective function. In this way, the independence of axes in the learned embedding space is no longer guaranteed. This limits the ability of InfoVGAE-SL to compute an informative representation and therefore reduces the F1 scores.

Effect of Rectified Gaussian Distribution: We apply a general Gaussian distribution instead of the rectified Gaussian distribution to learn the distribution of latent representations. Therefore, the values of latent variables become any real numbers rather than non-negative ones. We can observe from Table 5 that the performance of InfoVGAE-SL with the general Gaussian distribution is reduced.

Effect of KL Divergence Control: Next, we study the impact of the PI control algorithm on the performance of polarity detection. We remove the PI control algorithm in the InfoVGAE-SL. As illustrated in Table 5, its F1 scores for users and tweets decrease on all the datasets, especially for the Voteview dataset. This is attributed to the dense graph of Voteview dataset, since the voters usually vote for most of the bills. The dense graph leads to an unstable KL divergence during the training process. The KL control module helps control the KL divergence within a reasonable range, therefore the learned representations become more informative.

Effect of Sparsity Regularization: The InfoVGAE-SL without the sparsity regularization module is degrade to the previous InfoVGAE model [33]. By exploiting the sparsity regularization module, we observe an improvement on the F1 scores of users and tweets, as well as the clustering purity. This is because the sparsity regularization term helps enhance the sparsity of the learned belief representation and therefore results in a clearer separation of different ideologies.

Effect of Joint Learning of Tweets and Users: InfoVGAE-SL constructs BHIN containing both user and message nodes to jointly learn their embedding. We conduct an ablation study to test its effectiveness by separately building two graphs containing users or messages, and learning embeddings respectively. The evaluation metrics of the separate-learning version are 2%-6% lower than joint learning.

Based on the above ablation studies, we conclude that the total correlation module, non-negative latent space, PI control algorithm, sparsity regularization, as well as joint learning with BHIN plays an important role to learn a meaningful and disentangled embedding for polarity detection.

4.10 Ablation Study on Heterogeneous Content

The proposed InfoVGAE-SL model learns the ideological embedding based on heterogenous graph structure, it is naturally agnostic to content types. Therefore, it can handle multilingual, multi-platform, and multi-modal social messages (e.g. images, URLs, Hashtags). In this section, we experiment with message heterogeneity, by simultaneously considering different types of posts in the same source/message graph.

We augmented the original graph $\mathcal{G}$ (that captured Twitter posts and their sources) with two additional types of source/message pairs. The first represented online portals (such as news media) as source nodes and referenced URLs (e.g., article URLs) as the message nodes. The same URL address referenced by different users is grouped as one message. The second grouped similar images into visual clusters with users who posted each image representing sources and the image clusters representing (visual) messages. In this paper, we apply the same keypoint-based image clustering approach in [34] to cluster similar images together into the visual messages. We then evaluate the additional improvement of the proposed InfoVGAE-SL model over the baselines after applying

Table 6. Evaluation on Multi-Media Russia Ukraine War Dataset. We report the comparison of proposed InfoVGAE-SL model with baselines in the tweet only context, as well as the improvement of InfoVGAE-SL after incorporating the heterogenous URL and visual messages.

Model Name	User F1	message F1	Purity
NMF	0.9659	0.6797	0.8976
BSMF	0.6264	0.4171	0.6899
GCN	0.7883	0.5536	0.7709
DeepWalk	0.7769	0.7320	0.7329
TIMME-Unsup	0.9322	-	0.8373
InfoVGAE	0.9749	0.9287	0.9713
InfoVGAE-SL	0.9757	0.9374	0.9737
InfoVGAE-SL-URL	0.9759	0.9346	0.9729
InfoVGAE-SL-Visual	**0.9782**	**0.9433**	**0.9757**
InfoVGAE-SL-Visual-URL	0.9760	0.9374	0.9740

different kinds of heterogenous messages, where (i) only the original user/tweet interaction data are included, (ii) web URLs referred by the users are used as additional messages, (iii) the images referred by the users are included, (iv) both the URL and image messages are included. The evaluation result on the multi-media Russia Ukraine War dataset with human-labeled tweet annotations is shown in Table 6. The proposed InfoVGAE-SL model outperforms the baseline models on the Russian Ukraine War dataset in the original user/tweet interaction case, meanwhile, we've also seen 0.25% and 0.59% additional improvement of F1 score of users and tweets for InfoVGAE-SL model *after adopting the heterogenous types of asssertions*. The best F1 score is achieved when we include the visual messages. The URL messages can also help improve the ideology separation of users, but the improvement after adopting the URL messages is not as significant as the visual ones. The proposed InfoVGAE-SL model can not only support heterogenous messages, but also make use of the data to gain further information and improve ideology separation.

5 RELATED WORK

In a *previous conference publication* [33], the authors suggested that by combining a rectified (non-negative) Gaussian distribution with an inner-product decoder, variational graph auto-encoders could produce a disentangled and interpretable embedding space, thereby enhancing various downstream applications. However, it lacks a formal definition or proof of interpretability. *In this paper*, we: (i) formally define the properties of interpretable ideological embeddings; (ii) propose and prove broad conditions under which general graph embedding learning models can produce such interpretable embeddings; (iii) based on our proposed theory, present a concrete algorithm featuring two novel modules—sparsity regularization and local observation compensation—to enhance the interpretability of learned embeddings; and (iv) extend the applications of learned embeddings to broader datasets and tasks, including scenarios with heterogeneous content.

Aside from the interpretability of the latent space, the work generally falls in the category of research on stance detection and polarity classification [5, 9, 16, 22, 26, 36, 40, 53]. While some stance detection relied on sentiment analysis [9], most studies framed the stance detection problem as a supervised classification problem [16, 22, 26, 30, 32, 36] or a transfer learning problem [40] that correlates text features and user network neighborhoods with stance [16, 53]. Multi-target stance prediction explored correlations between different stances (e.g., with respect to election candidates) [50]. Traditional machine learning algorithms [7, 8, 62], such as SVM, Random Forest, and Bayesian estimators, were used. For example, Da Silva et al. [15] developed an ensemble method that combines RF, SVM, and LR to improve the classification accuracy.

With advances in deep neural networks, recent work applied deep learning models to detect polarity by mapping people's systems of belief into a latent space [26, 43, 47, 58, 61, 67]. For example, Jiang et al. [26] developed a weakly supervised model, Retweet-BERT, to predict the polarity of users on Twitter based on network structures and content features. Xiao et al. [67] proposed a multitask multi-relational embedding model to predict users' ideology using graph convolutional networks (GCN). However, the problem of supervised learning approaches is that they require human-annotated data, which is costly and time-consuming.

To deal with this issue, some work adopted unsupervised learning models for stance and/or polarity detection [1, 21, 25, 53, 59]. Unsupervised solutions were developed for clustering users by stance or viewpoint [59]. For example, Jang et al. [25] proposed a probabilistic ranking model to detect the stance of users in two communities. In addition, researchers [59] developed a purely unsupervised Author Interaction Topic Viewpoint model (AITV) for polarity detection at the post and the discourse levels. A recent solution extend polarity separation to visual content [35]. However, these methods do not focus on belief embedding.

Generalizing from stance classification problems, some work explored ideology as a variable that changes within a range [8]. It was postulated that human stances on issues can be predicted from a low-dimensional model [3, 11, 17, 68]. Ange et al. [3] developed a semi-supervised deep learning model based on multi-modal data for polarity detection. After that, Darwish et al. [17] adopted an unsupervised stance detection model that maps users into a low dimensional space based on their similarity. These methods, however, are mostly focused on either user polarity or statement polarity, but usually not both jointly. Extending this view, we develop an unsupervised belief representation learning model that jointly learns the latent representations of users and messages in the same space, thereby improving user and message polarity identification. Importantly, we think of unsupervised belief representation learning as a *separable problem* from the downstream application task. Thus, we show (in the evaluation) how the same approach is trivially applied to stance detection, stance prediction, and polarity separation, among other ideology-related analysis tasks in polarized settings.

6 CONCLUSION

In this paper, we define condition and feasibility of the explainable ideological representation learning and propose an Information-Theoretic Variational Graph Auto-Encoder with Sparsity Regularization (InfoVGAE-SL) for explainable ideological representation learning in an unsupervised manner. It constructs a bipartite heterogeneous graph from the interaction data and jointly learns the belief embedding of both users and their messages in the same latent space. InfoVGAE-SL includes four modules to better disentangle the latent space and learned more informative representations for downstream tasks. It adopts the rectified Gaussian distribution to create an orthogonal latent space, which assigns the belief systems into axes. A KL divergence PI controller is applied to enhance the disentanglement and a total correlation regularizer is proposed to learn a series of statistically independent latent dimensions. The sparsity regularization is exploited to enhance the orthogonality of representations and separation of ideologies. We also further propose a local observation compensation technique to empirically eliminate the bias from the Echo Chamber phenomenon. Experimental results show that the proposed InfoVGAE-SL outperforms the existing unsupervised polarity detection methods, and achieves a highly comparable F1 score and purity result with semi-supervised methods.

REFERENCES

[1] Leman Akoglu. 2014. Quantifying political polarity based on bipartite opinion networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8.

[2] Md Tanvir Al Amin, Charu Aggarwal, Shuochao Yao, Tarek Abdelzaher, and Lance Kaplan. 2017. Unveiling polarization in social networks: A matrix factorization approach. In *INFOCOM*. IEEE, 1–9.

[3] Tato Ange, Nkambou Roger, Dufresne Aude, and Frasson Claude. 2018. Semi-supervised multimodal deep learning model for polarity detection in arguments. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[4] Miguel A Arcones and Evarist Gine. 1992. On the bootstrap of U and V statistics. *The Annals of Statistics* (1992), 655–674.

[5] Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Who falls for online political manipulation?. In *Companion Proceedings of The 2019 World Wide Web Conference*. 162–168.

[6] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science* 26, 10 (2015), 1531–1542.

[7] Pablo Barberá and Gonzalo Rivero. 2015. Understanding the political representativeness of Twitter users. *Social Science Computer Review* 33, 6 (2015), 712–729.

[8] Pablo Barberá and Gaurav Sood. 2015. Follow your ideology: Measuring media ideology on social networks. In *Annual Meeting of the European Political Science Association.*

[9] Adam Bermingham and Alan Smeaton. 2011. On using Twitter to monitor political sentiment and predict election results. In *SAAIP*. 2–10.

[10] Adam Boche, Jeffrey B Lewis, Aaron Rudkin, and Luke Sonnet. 2018. The new Voteview. com: preserving and continuing Keith Poole's infrastructure for scholars, students and observers of Congress. *Public Choice* 176, 1 (2018), 17–32.

[11] Petko Bogdanov, Michael Busch, Jeff Moehlis, Ambuj K Singh, and Boleslaw K Szymanski. 2013. The social media genome: Modeling individual topic-specific behavior in social media. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 236–242.

[12] Mark J Brandt and Willem WA Sleegers. 2021. Evaluating belief system networks as a theory of political belief system dynamics. *Personality and Social Psychology Review* 25, 2 (2021), 159–185.

[13] Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. 2019. Isolating Sources of Disentanglement in VAEs. *arXiv preprint arXiv:1802.04942* (2019).

[14] Gary W Cox and Keith T Poole. 2002. On measuring partisanship in roll-call voting: The US House of Representatives, 1877-1999. *American Journal of Political Science* (2002), 477–489.

[15] Nadia FF Da Silva, Eduardo R Hruschka, and Estevam R Hruschka Jr. 2014. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems* 66 (2014), 170–179.

[16] Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. Improved stance prediction in a user similarity feature space. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*. 145–148.

[17] Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 141–152.

[18] Massimiliano Demata, Michelangelo Conoscenti, Stavrakakis Yannis, et al. 2020. Riding the Populist Wave. Metaphors of Populism and Anti-Populism in the Daily Mail and The Guardian. *Iperstoria* 15 (2020), 8–35.

[19] Carlos Diaz Ruiz and Tomas Nilsson. 2023. Disinformation and Echo Chambers: How Disinformation Circulates on Social Media Through Identity-Driven Controversies. *Journal of Public Policy & Marketing* 42, 1 (2023), 18–35.

[20] Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. 2019. Auto-encoding total correlation explanation. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 1157–1166.

[21] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*. 913–922.

[22] Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *EMNLP*. 751–762.

[23] Shanto Iyengar and Kyu S Hahn. 2009. Red media, blue media: Evidence of ideological selectivity in media use. *Journal of communication* 59, 1 (2009), 19–39.

[24] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1113–1122.

[25] Myungha Jang and James Allan. 2018. Explaining controversy on social media via stance summarization. In *SIGIR*. 1221–1224.

[26] Julie Jiang, Xiang Ren, and Emilio Ferrara. 2021. Social media polarization and echo chambers: A case study of COVID-19. *arXiv preprint arXiv:2103.10979* (2021).

[27] Hyunjik Kim and Andriy Mnih. 2018. Disentangling by factorising. In *ICML*. PMLR, 2649–2658.

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

[29] Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* (2016).

[30] Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)* 53, 1 (2020), 1–37.

[31] Jeffrey B Lewis, Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet. 2019. Voteview: Congressional roll-call votes database. *See https://voteview. com/(accessed 27 July 2018)* (2019).

[32] Jinning Li, Yirui Gao, Xiaofeng Gao, Yan Shi, and Guihai Chen. 2019. SENTI2POP: sentiment-aware topic popularity prediction on social media. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1174–1179.

[33] Jinning Li, Huajie Shao, Dachun Sun, Ruijie Wang, Yuchen Yan, Jinyang Li, Shengzhong Liu, Hanghang Tong, and Tarek Abdelzaher. 2022. Unsupervised Belief Representation Learning with Information-Theoretic Variational Graph Auto-Encoders. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1728–1738.

[34] Xinyi Liu, Jinning Li, Dachun Sun, Ruijie Wang, and Tarek Abdelzaher. 2023. Political Internet Memes and Political Learning: an Experimental Approach. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*.

[35] Xinyi Liu, Jinning Li, Dachun Sun, Ruijie Wang, Tarek Abdelzaher, Matt Brown, Anthony Barricelli, Matthias Kirchner, and Arslan Basharat. 2023. Unsupervised Image Classification by Ideological Affiliation from User-Content Interaction Patterns. In *Second Workshop on Images in Online Political Communication*.

[36] Walid Magdy, Kareem Darwish, Norah Abokhodair, Afshin Rahimi, and Timothy Baldwin. 2016. # isisisnotislam or# deportallmuslims? Predicting unspoken views. In *Proceedings of the 8th ACM Conference on Web Science*. 95–106.

[37] Daniele Mantegazzi. 2021. The geography of political ideologies in Switzerland over time. *Spatial Economic Analysis* 16, 3 (2021), 378–396.

[38] Nolan M McCarty, Keith T Poole, and Howard Rosenthal. 1997. *Income redistribution and the realignment of American politics*. AEI press.

[39] IC Mogotsi. 2010. Christopher d. manning, prabhakar raghavan, and hinrich schütze: Introduction to information retrieval.

[40] Mitra Mohtarami, James Glass, and Preslav Nakov. 2019. Contrastive language adaptation for cross-lingual stance detection. *arXiv preprint arXiv:1910.02076* (2019).

[41] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. 2010. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory* 56, 11 (2010), 5847–5861.

[42] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *KDD*. 701–710.

[43] Ammar Rashed, Mucahid Kutlu, Kareem Darwish, Tamer Elsayed, and Cansın Bayrak. 2020. Embeddings-Based Clustering for Target Specific Stances: The Case of a Polarized Turkey. *arXiv preprint arXiv:2005.09649* (2020).

[44] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*. Springer, 593–607.

[45] Huajie Shao, Shuochao Yao, Andong Jing, Shengzhong Liu, Dongxin Liu, Tianshi Wang, Jinyang Li, Chaoqi Yang, Ruijie Wang, and Tarek Abdelzaher. 2020. Misinformation Detection and Adversarial Attack Cost Analysis in Directional Social Networks. In *ICCCN*.

[46] Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang, and Tarek Abdelzaher. 2020. Controlvae: Controllable variational autoencoder. In *ICML*. PMLR, 8655–8664.

[47] Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019. Tweet stance detection using an attention based neural ensemble model. In *NAACL*. 1868–1873.

[48] Youjia Sima and Jialin Han. 2022. Online Carnival and Offline Solitude:"Information Cocoon" Effect in the Age of Algorithms. In *2022 8th International Conference on Humanities and Social Science Research (ICHSSR 2022)*. Atlantis Press, 2461–2465.

[49] Bridget Smart, Joshua Watt, Sara Benedetti, Lewis Mitchell, and Matthew Roughan. 2022. # IStandWithPutin versus# IStandWithUkraine: The interaction of bots and humans in discussion of the Russia/Ukraine war. In *Social Informatics: 13th International Conference, SocInfo 2022, Glasgow, UK, October 19–21, 2022, Proceedings*. Springer, 34–53.

[50] Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 551–557.

[51] Nicholas D Socci, Daniel D Lee, and H Sebastian Seung. 1998. The rectified Gaussian distribution. *NIPS* (1998), 350–356.

[52] Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *ACL*. 116–125.

[53] Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. In *ACL*. 527–537.

[54] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. 2012. Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics* 64, 5 (2012), 1009–1044.

[55] Dachun Sun, Chaoqi Yang, Jinyang Li, Ruijie Wang, Shuochao Yao, Huajie Shao, Dongxin Liu, Shengzhong Liu, Tianshi Wang, and Tarek F. Abdelzaher. 2021. Computational Modeling of Hierarchically Polarized Groups by Structured Matrix Factorization. *Frontiers in Big Data* 4 (2021).

[56] Lili Sun, Xueyan Liu, Min Zhao, and Bo Yang. 2021. Interpretable Variational Graph Autoencoder with Noninformative Prior. *Future Internet* (2021).

[57] Yizhou Sun and Jiawei Han. 2012. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery* 3, 2 (2012), 1–159.

[58] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. 2017. Multiple instance detection network with online instance classifier refinement. In *CVPR*. 2843–2851.

[59] Amine Trabelsi and Osmar Zaiane. 2018. Unsupervised model for topic viewpoint discovery in online debates leveraging author interactions. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[60] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)* (2018).

[61] Muhammad Umer, Zainab Imtiaz, Saleem Ullah, Arif Mehmood, Gyu Sang Choi, and Byung-Won On. 2020. Fake news stance detection using deep learning architecture (cnn-lstm). *IEEE Access* 8 (2020), 156695–156706.

[62] Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring user political preferences from streaming communications. In *ACL*. 186–196.

[63] Ruijie Wang, Zheng Li, Danqing Zhang, Qingyu Yin, Tong Zhao, Bing Yin, and Tarek Abdelzaher. 2022. RETE: Retrieval-Enhanced Temporal Event Forecasting on Unified Query Product Evolutionary Graph. In *The Web Conference*.

[64] Ruijie Wang, Yuchen Yan, Jialu Wang, Yuting Jia, Ye Zhang, Weinan Zhang, and Xinbing Wang. 2018. AceKG: A Large-Scale Knowledge Graph for Academic Data Mining. In *CIKM*.

[65] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *The World Wide Web Conference*. 2022–2032.

[66] Satosi Watanabe. 1960. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development* 4, 1 (1960), 66–82.

[67] Zhiping Xiao, Weiping Song, Haoyan Xu, Zhicheng Ren, and Yizhou Sun. 2020. TIMME: Twitter Ideology-detection via Multi-task Multi-relational Embedding. In *KDD*. 2258–2268.

[68] Chaoqi Yang, Jinyang Li, Ruijie Wang, Shuochao Yao, Huajie Shao, Dongxin Liu, Shengzhong Liu, Tianshi Wang, and Tarek F Abdelzaher. 2020. Disentangling Overlapping Beliefs by Structured Matrix Factorization. *arXiv e-prints* (2020), arXiv–2002.

[69] Poshan Yu, Yuejia Liao, and Ramya Mahendran. 2022. Research on Social Media Advertising in China: Advertising Perspective of Social Media Influencers. In *Handbook of Research on Global Perspectives on International Advertising*. IGI Global, 88–122.

[70] Hamed Zamani, Mostafa Dehghani, W Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 2018. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 497–506.

[71] Shengjia Zhao, Jiaming Song, and Stefano Ermon. 2019. Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the aaai conference on artificial intelligence*. 5885–5892.