PRISM: A Framework for Multi-Level Modeling and Analysis of Polarization Knowledge

Demetris Paschalides George Pallis Marios D. Dikaiakos
Computer Science Department, University of Cyprus
{dpasch01, pallis, mdd}@ucy.ac.cy

***Abstract*—Polarization poses a growing threat to democratic discourse, public trust, and societal stability. To better understand its structure and evolution, we conceptualize polarization as a multi-level phenomenon spanning entities, groups, and topics. We present PRISM, a framework that models polarization using a typed, weighted, and directed structure known as the Polarization Knowledge Graph (PKG). PRISM introduces a suite of analytical methods for multi-level analysis: i) identifying key actors and categorizing them as protagonists or antagonists based on their contribution to conflict, ii) measuring group cohesiveness through ideological alignment and topic-level agreement, and iii) ranking topics by their polarization intensity. We validate PRISM through a case study on U.S. COVID-19 media discourse, uncovering polarization patterns that align with established findings and highlight the politicization of the pandemic.**

***Index Terms*—Polarization, Polarization Knowledge Graph, Polarizing Topics, Ideological Cohesiveness, Polarizing Entities**

I. Introduction

Polarization is recognized as a global threat with profound implications for democratic stability, public trust, and collective action [1], [2]. Beyond politics, it shapes responses to social justice, climate policy, and public health. For instance, during the COVID-19 pandemic, polarization significantly influenced vaccine uptake and adherence to safety measures, contributing to higher mortality rates [3], [4]. These pervasive effects underscore the need to better understand how polarization emerges and evolves across different societal levels.

Social scientists define polarization as a "*social process where a social or political group is segregated into two or more opposing sub-groups with conflicting beliefs*" [5]. In this view, polarization can be conceptualized as a multi-level phenomenon spanning the entity, group, and topic-levels. At the entity-level, individuals form beliefs shaped by interactions with others, resulting in positive, negative, or neutral relationships [6]. At the group-level, entities form fellowships, emergent clusters of individuals whose alignment reflects shared views and social identity mechanisms [7]. Such fellowships often clash with each other, forming fellowship dipoles. At the topic-level, differences in entity attitudes toward key topics express and deepen these divides [8].

Existing research predominantly examines polarization at a single-level. Most commonly, studies focus on the group-level, quantifying disagreement between online user communities on specific topics by building interaction networks around a predefined subject (e.g. abortion), then partitioning users into opposing groups to measure separation [9]–[14]. A smaller body of work addresses topic-level polarization [9], [11], [15], estimating divergence in language or stance between pre-identified groups. However, entity-level polarization remains largely underexplored, with few efforts extending beyond post-hoc interpretation [16]–[18]. Single-level methods offer valuable insights but often miss broader patterns, such as how key entities influence polarization or how group dynamics shift across topics. These limitations highlight the need for a multi-level perspective capable of modeling and analyzing polarization across entities, groups, and topics.

To address the limitations of single-level polarization analysis, we propose PRISM[1], a framework for modeling and analyzing polarization as a structured, multi-level phenomenon (see Figure 1). PRISM leverages our previously introduced Polarization Knowledge Graph (PKG) [9], a typed, directed, and weighted graph that encodes polarization-relevant knowledge through three primary node types: *Entity*, *Fellowship*, and *Topic*, and can be instantiated from diverse sources like social media, voting records, or news corpora. Building upon this foundation, PRISM significantly extends the PKG with a comprehensive suite of analytical methods tailored for multi-level polarization analysis at the entity, group, and topic-levels. We specifically contribute novel analytical methods designed for this multi-level analysis:

- **Entity-level**: Novel techniques for identifying key polarization actors and classifying them as protagonists or antagonists, based on their contribution to conflict. This is achieved through a novel signed semantic association (SA) metric that quantifies each entity's impact on polarization dynamics.
- **Group-level**: A dedicated cohesiveness metric for assessing intra-fellowship dynamics, distinguishing between ideological alignment (e.g. Left vs. Right) and attitudinal agreement on specific discussion topics.
- **Topic-level**: A novel topic polarization score designed to quantify polarization intensity. This score effectively measures inter-group disagreement across topics and aggregates and ranks topics based on their dipole divergence, thereby identifying the most contentious subjects.

We validate PRISM through a case study on U.S. COVID-19 media discourse, demonstrating its ability to uncover multi-level polarization dynamics. Our findings show that political actors and ideological divisions often overshadowed health

[1]Source code: https://github.com/dpasch01/polarlib/tree/main/polarlib/prism

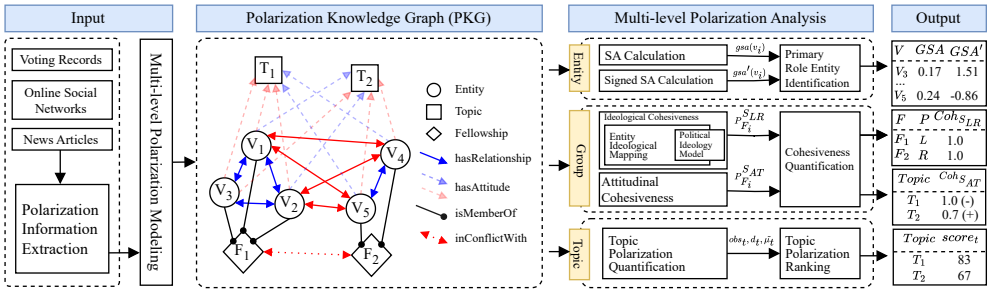

Fig. 1: Overview of the PRISM framework. PRISM constructs a PKG from input data sources, capturing entities, topics, fellowships, and their relationships (supportive (+) and oppositional (–) via the relevant predicates. PRISM then applies multi-level polarization analysis: at the entity-level, it computes global SA (GSA) and signed SA (GSA') to identify protagonists and antagonists; at the group-level, it measures ideological ($coh_{S_{LR}}$) and attitudinal fellowship cohesiveness ($coh_{S_{AT}}$); at the topic-level, it quantifies inter-group disagreement and ranks topics by their overall polarization score ($score_t$). The resulting outputs include entity role labels, group cohesiveness values, and a ranked list of polarizing topics.

authorities, highlighting the politicization of the pandemic. These results, aligned with literature [3], affirm PRISM's value for analyzing polarization in complex sociopolitical domains.

II. Related Work

Group-level Polarization: The majority of polarization studies examine how the phenomenon emerges within communities, such as social networks [12]–[14], political blogs [10], [14], [16], or voting records [19]. These group-level approaches typically model actor networks and quantify polarization based on the separation between manually or automatically defined partitions. For example, Adamic et al. [16] analyze the hyperlinks between the U.S. political blogs, measuring ideological clustering. Akoglu et al. [19] apply a signed bipartite graph model over congressional votes, introducing an unsupervised polarization metric. In social networks, Morales et al. 2015 [12] estimate user opinions in interaction networks, using a polarization index to capture disagreement.

Topic-level Polarization: Topic-level polarization methods analyze how ideological groups diverge in their discussions on specific issues, often using topic modeling [11], [20] or distributional semantics [15], [21]. Balasubramanyan et al. [20] introduce MCR-LDA, a model that captures emotional tone and topic prevalence across partisan communities. Demszky et al. [11] combine user ideology labeling with word embedding-based topic models to identify and compare political discourse on Twitter. More recent approaches like PaCTE [15] apply contextualized embeddings to quantify semantic distance between ideologically distinct corpora. These models surface salient divisive issues but rely on predefined user partitions and often treat group dynamics as static.

Entity-level Polarization: Entity-level polarization remains under-explored, as most studies emphasize group or topic-level dynamics. In the few cases where entities are examined, their roles are typically assessed post hoc, based on observed alignment or interaction patterns rather than through explicit modeling. For instance, Adamic et al. [16] analyze political blogs to identify partisan endorsements of public figures, and Akoglu et al. [19] infer ideological positioning from congressional voting records. Paschalides et al. [9] incorporate entities into a structured polarization model and evaluate static entity alignment based on attitude similarity. However, these approaches do not explicitly characterize the role of entities in shaping polarization dynamics. Consequently, existing works provide only limited insight into how individual actors contribute to the emergence and evolution of polarization.

Discussion: While some recent works explore multi-level polarization [9], [15], they focus on structural integration and lack targeted metrics to analyze how polarization emerges and interacts across levels. As a result, they offer limited insight into the dynamics or influence of entities, groups, and topics. PRISM addresses this gap by extending structural models with a suite of multi-level analytical methods. It quantifies the polarization roles of entities, measures intra-group alignment, and ranks topics by their extent of disagreement, thus offering an interpretable and operational understanding of polarization as an interconnected, multi-level process.

III. Multi-level Polarization Modeling

Our modeling framework is inspired by polarization social theory, where entities develop individual attitudes toward topics, influencing their relationships and leading to group-level polarization, which is manifested by fellowships and conflicting dipoles [7], [8]. At the topic level, these fellowships express opposing collective stances. To capture this computationally, we seek to i) identify entity roles, ii) analyze fellowship formation, and iii) quantify topic-level disagreement.

Polarization Knowledge Graph (PKG): The PKG is a heterogeneous, directed, and weighted graph [9], denoted as $G = (V, E)$, where each node $v \in V$ represents an actor in the polarization space and is typed as $\tau(v) \in$ *Entity*, *Fellowship*, *Topic*. The edges $e \in E$ encode a predicate $\lambda(e) \in \{$*hasRelationship*, *isMemberOf*, *hasAttitude*, *inConflictWith*$\}$, capturing relationships between actors (see Figure 2). A pair of *Entity* nodes v_i and v_j can be linked by a bi-directional predicate *hasRelationship*(v_i, v_j), which encodes observable interactions such as endorsements, replies, co-voting, or co-appearance in discourse. This edge is weighted by $w_{ij} \in [-1, 1]$, indicating their positive or negative interaction. Fellowships represent groups of entities that are densely

and positively connected, with each group represented by the *Fellowship* node type. Entities are linked to their fellowships through the predicate *isMemberOf*(v_i, v_k), where v_k is a *Fellowship* node. Fellowships may form dipoles when the inter-group connections between their members are predominantly negative. These conflicts are modeled with the predicate *inConflictWith*(v_k, v_z) between two *Fellowship* nodes, and the edge is weighted by $w_{kz} \in [0, 1]$, indicating the degree of polarization between them. *Topic* nodes represent subjects of discourse. Entities express their stance toward topics using the predicate *hasAttitude*(v_i, v_x), where $w_{ix} \in [-1, 1]$ reflects the entity's level of opposition (–1) or support (+1) for topic v_x.

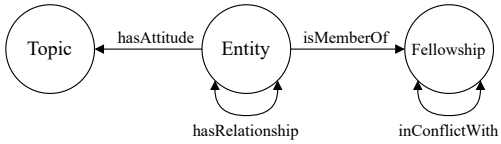

Fig. 2: Schema for the Polarization Knowledge Graph.

Entity-level Polarization: Entity-level polarization captures the diverse beliefs, attitudes, and interactions among individuals or organizations within a society [6]. In the PKG, these entities are modeled alongside fellowships and topics, connected through weighted predicates that reflect attitude, affiliation, and conflict [22]. Within entity-level polarization, a key challenge is identifying primary entities that drive polarization by occupying central positions in the narrative landscape [9], [11], [18], [19]. These entities exhibit strong semantic associations (SA) [23] with others, indicating narrative proximity and influence. We extend SA to account for signed relationships in the PKG, introducing a signed semantic association metric that reflects both connection strength and attitude (see Section IV). Using this, we classify entities as protagonists, who foster positive relations, or antagonists, who reinforce division and conflict. The metric ranges from –1 to 1 and is grounded in structural balance theory [24], enabling a principled assessment of each entity's polarization role.

Group-level Polarization: At the group level, polarization is fundamentally characterized by the emergence of distinct fellowships within a larger social structure [5]. In the context of the PKG, a fellowship is defined as a cluster of entities sharing common characteristics, such as ideological alignment and similar attitudes toward topics. When such fellowships diverge in these dimensions, they form fellowship dipoles [7]. The formation and interpretability of fellowships depend on their internal cohesiveness, which reflects the consistency of member attitudes and ideologies. High cohesiveness reduces the likelihood of intra-group conflict, making inter-fellowship (i.e. group-level) polarization more meaningful and reliable [8]. Conversely, distinct separation between cohesive fellowships signals stronger polarization between them [10]. To assess group-level polarization, we examine two core aspects of fellowship members: i) their ideologies (e.g. Left vs. Right), and ii) their attitudes (support or opposition) toward various topics. These dimensions define two types of cohesiveness: ideological cohesiveness, which reflects alignment in political orientation [10], [25], and attitudinal cohesiveness, which captures agreement on topic-specific stances [15]. In Section V we introduce metric functions to quantify both forms of cohesiveness within fellowships and use them to assess the degree of group-level polarization.

Topic-level Polarization: Topic-level polarization captures divisions in discourse by assessing how entities express opposing attitudes toward specific topics. Within the PKG, these are represented via the predicate *hasAttitude*, connecting *Entity* nodes to *Topic* nodes, each weighted by sentiment attitude. Measuring topic-level polarization requires: i) quantifying disagreement in attitudes for a given topic, and ii) ranking topics by their overall polarization intensity. We distinguish between local and global topic polarization, with the former measuring disagreement on a topic within a single fellowship dipole, and the latter aggregating disagreement across all dipoles, offering a broader view of how polarizing a topic is in the PKG. To estimate local polarization for each topic within individual dipoles, we apply the polarization index metric [12], which quantifies the dispersion of entity attitudes (e.g. support vs. opposition), yielding a score from 0 (no polarization) to 1 (extreme polarization). For the global topic polarization computation, we propose a ranking function that aggregates local polarization scores while accounting for both the intensity of disagreement and the prominence of each topic across dipoles. More details are provided in Section VI.

IV. Primary Polarization Entities

In the PKG, the SA between a pair of *Entity* nodes captures their contextual relatedness within the polarization narrative. We define this as **local SA**, estimated by the overlap in their neighboring entity nodes (connected via *hasRelationship* predicate), using the Normalized Google Distance [23], a standard metric for computing SA in structured information networks. Given two entities $a, b \in V$ such that $\tau(a) = \tau(b) =$ *Entity*, their local SA is computed as:

$$lsa(a,b) = \begin{cases} \frac{log(max(|A|,|B|)) - log(|A \cap B|)}{log(|V|) - log(min(|A|,|B|))}, & \text{if } |A \cap B| > 0 \\ 0, & \text{otherwise} \end{cases}$$

where A and B denote the sets of neighboring entities adjacent to a and b, respectively. Entities with high degree and large overlap in neighbors yield higher local SA values.

To identify primary polarization entities, we define the **global SA** of an entity v_i, denoted as $gsa(v_i)$. This is computed by summing its **local SA** scores with all neighboring entities within the PKG. The local SA score quantifies the semantic relatedness between two specific entities. Therefore, global SA reflects the overall extent to which an entity is semantically embedded across the PKG, with higher global SA indicating a more prominent or influential role in the polarization narrative.

A. Protagonists and Antagonists

To examine the roles that entities play in the polarization landscape, we distinguish between two categories: *protagonists* and *antagonists*. Protagonists are entities that contribute

to balancing or stabilizing polarization, potentially mitigating its effects by fostering positive relationships. In contrast, antagonists intensify polarization, often by reinforcing negative relations or fueling inter-group conflict. To identify these roles, we extend SA to a signed version that incorporates the polarity of *hasRelationship* edges.

Given that a *hasRelationship* predicate connects entities a and b, and $A \cap B$ denotes their common neighbors, we assert that for each $c \in A \cap B$, a triangle exists in the PKG among a, b, and c. Due to the signed nature of the *hasRelationship* predicate, these triangles exhibit polarity. According to structural balance theory [24], a triangle with three positive edges reflects the notion that "*the friend of my friend is my friend*", whereas those with one positive and two negative edges express variations such as "*the friend of my enemy is my enemy*", "*the enemy of my friend is my enemy*", and "*the enemy of my enemy is my friend*". To quantify these notions, we identify the intersections of $A^+ \cap B^+$, $A^+ \cap B^-$, $A^- \cap B^+$, and $A^- \cap B^-$, where A^+, A^-, B^+, and B^- correspond to the positively (friends) and negatively (enemies) related neighbors of a and b respectively. We interpret $A^+ \cap B^+$ as indicating a positive association between a and b through their common positive connections. In the context of structural balance, the only permissible case where $|A^+ \cap B^+| > 0$ is when the a and b sign is positive . The resulting sets of $A^+ \cap B^-$ and $A^- \cap B^+$ denote an overall negative association, indicating that an entity's friends are the enemies of the other, and vice versa . Finally, $A^- \cap B^-$ denotes a positive association and reflects that the enemies of a are also the enemies of b. This is only applicable if the connection of a and b is positive . Consequently, we define the **signed local SA** between entities a and b, based on their neighborhoods A and B, as:

$$lsa'(A,B) = \begin{cases} lsa(A^+,B^+) + lsa(A^-,B^-), & \text{if } w_{a,b} > 0 \\ lsa(A^+,B^-) + lsa(A^-,B^+), & \text{if } w_{a,b} < 0 \\ 0, & \text{otherwise} \end{cases}$$

where $w_{a,b}$ represents the sign (± 1) of the *hasRelationship* edge between a and b. The pairwise local signed SA scores populate a $|V| \times |V|$ matrix M'. The **signed global SA** for an entity is computed as the sum of its neighboring local signed SA as $gsa'(v_i) = \sum_{j=1}^{|V|} M'_{ij}$, $i \neq j$.

V. Fellowship Cohesiveness

Group cohesiveness is a key indicator of polarization in social and political contexts [8]. In the PKG, fellowships represent clusters of aligned entities, and their cohesiveness reflects the degree of internal agreement. To assess polarization, we introduce metrics that quantify the cohesiveness of fellowships based on shared properties and attitudes among their members, determined via the *isMemberOf* predicate. Cohesiveness may be *ideological*, capturing alignment along a political spectrum, or *attitudinal*, measuring agreement on specific topics (e.g. abortion, gun control). High ideological cohesiveness signifies a well-defined segmentation and is associated with increased polarization [10], [25], while attitudinal cohesiveness reveals intra-group consensus or contention [8].

To operationalize this, we model an entity's position on an ideology or topic using a discrete value within a completely ordered, finite set, representing an *ideological or attitudinal spectrum*. For ideological cohesiveness, we use a fixed spectrum such as $S_{LR} = \langle \textit{Left}, \textit{Moderate}, \textit{Right} \rangle$, with positions typically assigned using external knowledge (e.g. known political affiliation). For attitudinal cohesiveness, we define topic-specific spectra, such as $S_{CV} = \langle \textit{OppositionCV}, \textit{SupportCV} \rangle$ for the COVID-19 Vaccines topic. Entities are mapped to these spectra based on the weights of their *hasAttitude* predicates in the PKG, with supportive or oppositional positions determined via thresholds that may be specified manually, or inferred from the distribution of attitude scores. Following, we define the cohesiveness metric, and outline entity mapping onto ideological and attitudinal spectra.

A. Fellowship Cohesiveness Metric

For fellowship F and spectrum S, we define cohesiveness as:

$$coh_S(P_F^S) = \tfrac{1}{N} \times \textstyle\sum_{p_k \in P_F^S} d(p_k, z)$$

where P_F^S is the defined mapping of entities in F to positions on spectrum S. z corresponds to the modal spectrum position of the entities in F, and N the number of entities in F. The similarity function $d(s_i, s_j) = 1.0 - q/n$ computes the closeness that two positions s_i and s_j manifest on the ordered spectrum S, where q is the number of consecutive steps between them and n is the total number of positions in S. The proposed metric outputs values in $[0, 1]$, with 1 indicating perfect cohesiveness (i.e. all members share the same position) and 0 indicating maximal dispersion. Because spectrum positions reflect meaningful distinctions on an ordinal scale, closer positions are considered more similar (e.g. position i is more similar to $i+1$ than $i+2$). The similarity function d captures this proximity-based similarity. This metric assesses how well aligned the fellowship members are by averaging their similarity to the modal position, thus quantifying intra-group cohesiveness while respecting the structure of the underlying ideological or attitudinal spectrum.

B. Ideological Cohesiveness

We compute the ideological cohesiveness of fellowship F as $coh_{S_{LR}}(P_F^{S_{LR}})$ where the S_{LR} denotes the ideological spectrum, and $P_F^{S_{LR}}$ is the mapping of fellowship entities to spectrum positions. Constructing this mapping requires an external knowledge source K that provides political affiliation data. Formally, the ideological mapping is defined as a function $f_{S_{LR}}(F, K) \rightarrow S_{LR}$, which assigns spectrum positions to the entities in F based on knowledge from K.

In practice, obtaining a knowledge source K that covers all entities in a PKG is challenging. This task is more tractable in domains with strong political relevance, such as gun control or election discourse. To address these gaps in K, we adopt a weakly supervised approach. We begin with a set of seed political entities (PEs), with known ideological positions, and then propagate this information to non-political entities (NPEs)

by leveraging the structure of the PKG, particularly the signed weights of the *hasRelationship* predicates. This propagation allows us to infer the likely ideological stance of entities indirectly involved in political discourse.

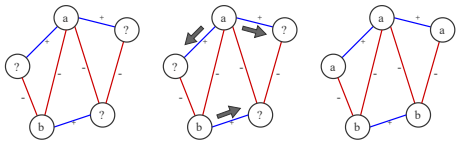

Fig. 3: Step-by-step application of WLPA on SAG.

Propagating Political Positions: Given a small set of seed political entities (PEs) with known ideological positions, we infer the positions of non-political entities (NPEs) through propagation. To do this, we apply the Weighted Label Propagation Algorithm (WLPA) [26], leveraging the signed nature of the *hasRelationship* predicate. As illustrated in Figure 3, positively weighted (supportive) relationships increase the likelihood of ideological alignment, while negatively weighted (oppositional) relationships reduce it. During each WLPA iteration, the political position of an entity $v_i \in$ NPE is updated via weighted majority voting among its neighbors, with each vote scaled by the strength and sign of its connection to v_i. The algorithm proceeds for k iterations or until convergence. Notably, the approach generalizes beyond political contexts and supports other ideological spectra defined in the PKG.

C. Attitudinal Cohesiveness

Attitudinal cohesiveness reflects the degree of agreement among fellowship members in their supportive or oppositional attitudes toward a given topic. To compute this in PRISM, we map each entity's attitude toward a topic t_k to an attitudinal spectrum $S_{AT} = \langle \textit{Opposition}_{t_k}, \textit{Support}_{t_k} \rangle$. Each attitude is quantified via the *hasAttitude* predicate weight $w_{jk} \in [-1, 1]$. To discretize attitudes we use the mapping function ϕ:

$$\phi(w_{jk}) = \begin{cases} \textit{Opposition} & \text{if } w_{jk} \leq thr_{OPP} \\ \textit{Support} & \text{if } w_{jk} \geq thr_{SUP} \\ \textit{Neutral} & \text{otherwise} \end{cases}$$

where thr_{OPP} and thr_{SUP} are user-defined thresholds tailored to each case study. The resulting mapping for a fellowship F toward topic t_k is denoted $P_F^{S_{AT}^{t_k}} = \{\phi(w_{zk}) \mid v_z \in F, w_{zk} \neq 0\}$. Neutral attitudes are excluded to focus on entities taking explicit positions. The attitudinal cohesiveness of F toward t_k is then computed as $coh_{S_{AT}}(P_F^{S_{AT}^{t_k}})$, capturing how aligned the fellowship is on the topic.

VI. Topic Polarization Ranking

Topics play a central role in understanding polarization, serving as anchors for both its local and global manifestations. Local topic polarization captures the degree of attitudinal divergence between specific fellowship dipoles on a given topic, while global topic polarization reflects the overall divisiveness of a topic across the entire domain. To assess local polarization in PRISM, we apply the polarization index [12], which quantifies disagreement in attitudes within individual dipoles for a specific topic. To estimate global polarization, we introduce an aggregation function that combines the number of dipoles engaged with the topic, the local intensity of polarization within those dipoles, and the volume of attitude observations expressed toward the topic. Together, these measures offer a multi-faceted view of how topics contribute to polarization at different granularity levels.

Local Topic Polarization: A dipole in the PKG, denoted as D_{ij}, is represented by the predicate *inConflictWith* between fellowships F_i and F_j. For each topic t discussed by entities in these fellowships, we compute a local topic polarization score based on their expressed attitudes. Let A_t be the set of attitude weights w_{xt} associated with entities $v_x \in F_i \cup F_j$ toward topic t, as encoded in the PKG via the *hasAttitude* predicate. We partition A_t into A^+ and A^-, containing positive (supportive) and negative (oppositional) attitudes, respectively. The local polarization of topic t for dipole D_{ij} is then calculated using the polarization index $\mu = (1 - \Delta_A)\delta_A$, where Δ_A is the normalized difference in the sizes of A^+ and A^-, and $\delta_A = |gc^+ - gc^-|/2$ measures the average attitudinal divergence between the two sides. Here, gc^+ and gc^- denote the mean attitude values in A^+ and A^-, respectively. μ ranges from 0 (no polarization) to 1 (extreme polarization), capturing the local disagreement over topic t within dipole D_{ij}.

Global Topic Polarization: In PRISM, we rank topics by assessing both attitudinal disagreement and discussion coverage across fellowship dipoles. For each topic t, a vector of local polarization indices μ is computed, each derived from a dipole where t is discussed. To quantify overall topic polarization, we compute the median of these local indices, denoted $\tilde{\mu}_t$, reflecting the central tendency of disagreement across dipoles. However, relying solely on $\tilde{\mu}_t$ may bias rankings toward rarely discussed topics with extreme disagreement, overlooking widely discussed yet consistently polarized topics. To address this, we incorporate a scoring function that considers both the extent of discussion and the strength of polarization:

$$score_t = \left(\frac{obs_t}{d_t}\right) \cdot \tilde{\mu}_t$$

where obs_t is the number of *hasAttitude* observations related to topic t, and d_t is the number of dipoles in which t is discussed. The ratio $\frac{obs_t}{d_t}$ reflects the average engagement per dipole, modulating the median polarization $\tilde{\mu}_t$ to penalize sparse or unreliable cases. As a result, topics with broad discussion and consistent polarization rank higher, whereas those with isolated disagreement but limited coverage are penalized.

VII. Polarization amidst COVID-19 Pandemic

In early 2020, COVID-19 rapidly spread worldwide, leading to over 6M infections and 400K deaths within a year, many in the U.S. Research suggests that partisan polarization, amplified by mass media, significantly shaped public response and worsened health outcomes in the U.S. [4], [15], [27]. We apply PRISM to analyze polarization in U.S. news media during the COVID-19 pandemic [15] by automatically constructing a focused PKG using an existing framework, namely POLAR [9].

Our multi-level analysis includes: i) identifying key actors via global and signed global SA; ii) evaluating ideological and attitudinal fellowship cohesiveness, and iii) ranking the most polarizing topics based on local and global disagreement. Findings are contextualized against prior literature [4].

A. Computing the Polarization Knowledge Graph

To investigate polarization during the COVID-19 pandemic, we construct a PKG using the POLAR framework [9] applied to an existing dataset of approximately 66K U.S. news articles published between January and July 2020 [15]. POLAR extracts entities, relationships, fellowships, dipoles, topics, and attitudes using named entity recognition and linking, sentiment attitude analysis, and signed network clustering. Within this process, conflicts between fellowships (*inConflictWith* edges) emerge directly from the aggregation of negative entity-to-entity ties, with their weights reflecting the normalized intensity of these antagonistic cross-group connections. The resulting PKG integrates these structural and content-based elements, capturing how entities relate to one another and to key discussion topics within the pandemic narrative.

The constructed PKG includes 145 entities, 493 relationships, 45 fellowships, 104 dipoles, 101 topics, and over 10K attitude observations. Political figures dominate the entity landscape. Donald Trump appears over 34K times, followed by China, Joe Biden, and others, far outnumbering mentions of health authorities such as the WHO and CDC, reflecting the politicization of the crisis [4]. Most fellowships are small, with 34 out of 45 consisting of a single entity, but the top four largest account for over half the entities in the PKG, revealing a clear partisan divide: one group clusters around conservatives like Trump and Ted Cruz, while another centers on liberals like Biden and Kamala Harris. On the topic front, polarization was initially captured across approximately 1,000 clusters of noun phrases. To streamline analysis, human annotators labeled and consolidated these into 101 high-level discussion topics with strong inter-rater agreement (0.83), enabling structured assessment polarization during the pandemic.

B. Computing the Ideological Cohesiveness

To measure ideological cohesiveness, we define a mapping function $f_{S_{LR}}(F, K) \rightarrow S_{LR}$ that assigns positions on the ideological spectrum S_{LR} to fellowship members F, using an external knowledge resource K. As the PKG is built from a news corpus using NERL methods, we adopt Wikipedia as K, due to its compatibility with NERL outputs and its rich, structured political information [28]. Specifically, we determine the political position of entities via Wikipedia's Infobox tab, which includes relevant fields such as *Political Party*, *Political Position*, and *Ideology*, with the latter being most frequently populated. For each political entity $v \in$ PE, we extract its ideology context $ideo_v = \{u_1, ..., u_n\}$ as a set of textual tags from the *Ideology* field, and use it to estimate the entity's position on S_{LR} via a classifier $\theta(ideo_v) \rightarrow S_{LR}$.

To implement the estimation function θ, we fine-tune BERT [29], [30], a transformer-based language model, for ideological classification. BERT is well-suited to this task due to its ability to capture semantic relationships between ideological tags. Given a set of tags $ideo_v$, BERT tokenizes the input and generates contextual embeddings through its 12-layer encoder with 768-dimensional hidden states. The resulting representation is passed through a softmax classifier to produce a probability distribution over the positions in S_{LR}: $p(c|h) = \text{softmax}(h)$. We train the model on a labeled dataset of entities and their ideological positions [28], using standard hyperparameters and 5-fold cross-validation. The model, visualized in Figure 4, achieves an average F1 score of 93%, demonstrating robust performance. Once ideological positions for seed PEs are inferred, they are propagated across the PKG using the method described in Section V.

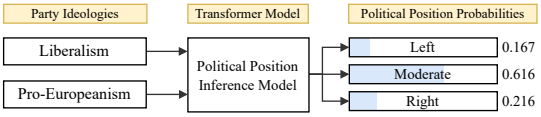

Fig. 4: Diagram of political position inference model.

VIII. Experiments and Discussion

A. Entities with Primary Polarization Roles

To analyze polarization at the entity-level, we use the global and signed global Semantic Association scores (GSA and GSA'). GSA highlights entities that played primary roles in the pandemic discourse, while GSA' reveals whether their associations were predominantly positive (protagonistic) or negative (antagonistic). As shown in Table I, Donald Trump and China rank highest in GSA (0.96 and 0.90), followed by the Democratic Party (0.25) and Joe Biden (0.22). The sharp drop (≈0.65) after Trump and China suggests their central presence and dense connectivity in the PKG.

We observe that Trump and China also top the list of antagonists with the most negative GSA' scores (–0.435 and –0.222), indicating they were involved in largely conflictual relations. Conversely, the Republican and Democratic Committees, Joe Biden, and Bernie Sanders appear as protagonists with positive GSA' scores. These patterns reveal that while Trump and China dominated the narrative, they often did so in ways that intensified polarization. Meanwhile, traditional political organizations and candidates maintained more cooperative or stabilizing positions. These trends reinforce prior findings about the political framing of the pandemic in U.S. media [4] and contrast with health authorities like the WHO (GSA=0.043, GSA'=-0.035), CDC (GSA=0.035, GSA'=0.010), and Anthony Fauci (GSA=0.028, GSA'=-0.028), who appeared less frequently and with weaker associative scores.

	Global SA		Top Positive GSA'		Top Negative GSA'	
No.	**Entity**	GSA	**Entity**	GSA'	**Entity**	GSA'
1	Donald Trump	0.96	Rep. Committee	0.035	Donald Trump	-0.435
2	China	0.90	Joe Biden	0.026	China	-0.222
3	Democratic Party	0.25	Dem. Committee	0.026	Republican Party	-0.141
4	Joe Biden	0.22	Bernie Sanders	0.023	U.S. Senate	-0.098
5	Republican Party	0.18	Jeff Sessions	0.020	Russia	-0.089
6	White House	0.14	Amy Klobuchar	0.020	U.S. House of Reps.	-0.072
7	U.S. Senate	0.11	U.S. Navy	0.019	Nancy Pelosi	-0.070
8	Russia	0.10	Andrew Cuomo	0.018	Mitch McConnell	-0.059

TABLE I: Entities with most GSA, and most / least GSA'.

Takeaway: Polarization was driven primarily by political entities. Donald Trump and China were the most prominent antagonists, exhibiting high number of conflictual ties. In contrast, figures like Joe Biden and party committees acted as protagonists, fostering more positive associations. Health authorities, such as the WHO, CDC, and Anthony Fauci, were notably less central, both in connectivity and polarization.

B. Ideological Cohesiveness

We assess the ideological cohesiveness of the four largest fellowships identified in the PKG (Table II). Fellowship 1 is mostly Republican (≈80%), except for two Democrats, namely Gavin Newsom and Roy Cooper, who exhibit positive ties to Donald Trump. Fellowship 2 includes primarily Democrats, with Joe Biden at its center. Fellowship 3 leans Left, with 5 of 7 members labeled as such. Although the CDC and FDA lack explicit ideological labels, their positive ties to Democratic actors suggest a Left-leaning alignment, likely reflecting the political climate during the pandemic, rather than a fixed partisan stance. Fellowship 4 combines Republicans like Mike Pence with institutional entities such as the White House, FBI, and House of Reps., many of which are politically unassigned.

Fellowship 1	***P***	**Fellowship 2**	***P***	**Fellowship 3**	***P***	**Fellowship 4**	***P***
Donald Trump	R	Joe Biden	L	Andrew Cuomo	L	Mike Pence	R
Tom Cotton	R	Democratic Party	L	U.S. Navy	L	Kayleigh McEnany	R
Ted Cruz	R	Bernie Sanders	L	Ron DeSantis	R	Raymond Flynn	L
Roy Cooper	L	Barack Obama	L	CDC	L	FBI	-
Gavin Newsom	L	Hillary Clinton	L	China	-	U.S. HR	-
Rep. Committee	R	Dem. Committee	L	Xi Jinping	L	White House	-
Mark Cuban	R	Kamala Harris	L	FDA	L		
Mark Meadows	R	PPACA	L				
Planned Parenthood	L	Elizabeth Warren	L				
HHS	R	U.S. Congress	L				
Howard Stern	R	U.S. Secret Service	R				

TABLE II: Fellowships with their infered political positions.

The ideological cohesiveness scores $coh_{S_{LR}}$ for fellowships 1 to 4 are 0.67, 0.93, 0.71, and 0.50, respectively. These relatively high values, despite the large fellowship sizes, indicate strong internal ideological agreement. This alignment reinforces the view that COVID-19 coverage reflected a politically polarized structure, with clear partisan entity clustering.

Takeaway: The largest fellowships divide along party lines, with high ideological cohesiveness confirming the presence of politicization. While some entities, like the CDC and FDA, are not inherently political, their inferred positions highlight how affiliations during the pandemic were politically aligned.

C. Attitudinal Cohesiveness

To assess attitudinal cohesiveness, we mapped each entity's topic-level attitude using empirically determined thresholds ($thr_{OPP} = -0.1$, $thr_{SUP} = 0.1$) and focused on the two largest fellowships. Fellowship 1 includes attitudinal observations for 95 topics, while fellowship 2 covers 88. Table III presents $coh_{S_{AT}}$ scores for selected topics. Both fellowships show high cohesiveness on topics such as the Coronavirus Stimulus Package, Chinese Propaganda, Economic Crisis, Reopening Plan, and Health Care Workers.

Despite topical overlap, the direction of attitudes differs between fellowships. Fellowship 1 consistently supports economic reopening and opposes scientific institutions (e.g. Medical Experts, Coronavirus Task Force), reflecting a conservative stance aligned with prior studies [3], [4]. In contrast, fellowship 2 supports public health efforts and criticizes Trump's pandemic response and re-election campaign, aligning with liberal positions. These patterns align with the ideological cohesiveness scores and reinforce partisan divisions on pandemic-related issues.

Topic	**Fellowship 1**			**Fellowship 2**		
	SUP	**OPP**	$coh_{S_{AT}}$	**SUP**	**OPP**	$coh_{S_{AT}}$
Coronavirus Stimulus Package	10	1	**0.91**	8	3	**0.73**
Trump Coronavirus Response	5	6	0.55	2	9	**0.82**
Trump Re-election Campaign	6	5	0.55	3	8	**0.73**
Joe Biden Campaign	7	4	0.64	7	4	0.64
Chinese Propaganda	0	11	**1.00**	1	10	**0.91**
Protests	5	6	0.55	4	7	0.64
Economic Crisis	8	3	**0.73**	10	1	**0.91**
Travel Restrictions	7	4	0.64	0	11	**1.00**
Medical Experts	2	9	**0.82**	8	3	**0.73**
Mask Mandate	6	5	0.55	7	4	0.64
Coronavirus Restrictions	5	6	0.55	7	4	0.64
Lockdown	2	9	**0.82**	4	7	0.64
Coronavirus Task Force	3	8	**0.73**	11	0	**1.00**
Immigration	4	7	0.64	10	1	**0.91**
Unemployment	9	2	**0.82**	6	5	0.55
Reopening Plan	10	1	**0.91**	2	9	**0.82**
Health Care Worker	10	1	**0.91**	8	3	**0.73**

TABLE III: Attitudinal cohesiveness measures ($coh_{S_{AT}}$) for fellowships 1 and 2 by topic.

Takeaway: Fellowships 1 (Right-leaning) and 2 (Left-leaning) display strong internal agreement on numerous topics. Fellowship 1 supports reopening and economic measures while opposing medical guidance, whereas fellowship 2 supports health interventions and criticizes political responses. Both show unified stances on shared concerns like economic relief and Chinese Propaganda, reflecting consistent ideological divides in pandemic discourse [3].

D. Polarizing Topic Ranking

To understand how polarization concentrates around specific issues, we analyze attitudes toward the 101 annotated topics captured in the PKG. For each topic, we retrieve the attitudes from the 104 dipoles identified in the PKG and compute local polarization using the polarization index [12]. We then aggregate these values using the function $score_t$ to estimate the topic's global polarization level. Table IV presents the top 10 most polarizing topics based on these scores.

No.	Topic t	d_t	obs_t	$\tilde{\mu_t}$	From	To	$score_t$
1	Elections	44	2,769	0.90	0.84	0.96	56.71
2	Stimulus Package	65	4,357	0.73	0.69	0.77	49.10
3	COVID-19 Cases	82	4,023	0.86	0.82	0.90	42.28
4	Trump Response	49	2,782	0.68	0.47	0.80	38.83
5	Trump Campaign	43	1,985	0.76	0.69	0.83	35.17
6	Press Briefings	36	1,187	0.78	0.69	0.87	25.61
7	Restrictions	25	688	0.89	0.86	0.91	24.41
8	Medical Experts	41	1,322	0.73	0.68	0.78	23.65
9	Protests	48	1,253	0.86	0.78	0.94	22.38
10	Mask Mandate	51	1,391	0.73	0.68	0.79	20.00

TABLE IV: Rank list of top-10 polarizing topics.

The 2020 U.S. Elections emerge as the most polarizing topic ($\tilde{\mu}$ = 0.90), with 2,769 observations across 44 dipoles. It is followed by the Coronavirus Stimulus Package ($\tilde{\mu}$ = 0.73) and Trump's Coronavirus Response ($\tilde{\mu}$ = 0.68), with the latter discussed in 49 dipoles and tied to 2,782 attitude observations. Coronavirus Restrictions also rank highly, with a $\tilde{\mu}$ of 0.86 across 25 dipoles. Below, we qualitatively examine selected cases to interpret their polarization scores.

Coronavirus Stimulus Package: although presented as a bipartisan measure, it exposes conflict over legislative support, reflected in contrasts such as "approve" and "block" (see Figures 5d and 5a). Most Democrats and Republicans backed the bill, though some, like Republican Thomas Massie, opposed it. The discourse also featured politically charged proposals, including legislation to permit lawsuits against China.

Trump Coronavirus Response: has been a controversial topic with observations that were both "*criticized*" and "*praised*". This topic illustrates intra-fellowship polarization, as Right-leaning entities in fellowship 1, diverge, with some supporting Trump's decisions while others place blame [3]. This is reflected in the group's moderate attitudinal cohesiveness (0.55), indicating substantial internal disagreement.

Restrictions: consists of polarizing partisan attitudes toward lockdowns, social distancing, and mask mandates. Democrats generally support these measures, while Republicans often reject them as "*strict*", suggesting re-opening.

(a) CARES Act (P) (b) Trump Resp. (P) (c) Restrictions (P)

(d) CARES Act (N) (e) Trump Resp. (N) (f) Restrictions (N)

Fig. 5: Positive (P) and negative (N) topical wordclouds.

Takeaway: Topics such as the 2020 Elections, Trump's pandemic response, the stimulus package, and public health restrictions show the highest levels of polarization in the PKG. These patterns reveal sharp and persistent ideological divides, particularly between Democratic and Republican fellowships.

IX. Conclusion and Future Work

In this work, we introduced PRISM, a framework for modeling and analyzing polarization at the entity, group, and topic-levels. PRISM identifies key polarization actors, measures fellowship cohesiveness, and ranks topics by polarization intensity. Applied to U.S. COVID-19 discourse, it revealed polarization dynamics consistent with established political science findings, underscoring its effectiveness in capturing complex sociopolitical patterns. While our case study focused on political ideology during the pandemic, broader evaluation remains an important direction. Future work will extend PRISM's assessment through additional case studies and systematic comparisons with established polarization metrics, while also examining scalability and complexity to ensure robustness. Although demonstrated in a political context, PRISM's definitions and metrics are spectrum-agnostic, enabling application to cultural, identity-based, or belief-driven dimensions. Validating the framework across these varied domains, and supporting custom spectrum definitions, will further highlight its generalizability and impact.

Acknowledgments: This research is funded in part by the EU Commission via the ATHENA 101132686 project (HORIZON-CL2-2023-DEMOCRACY-01).

References

[1] S. Aral and D. Eckles, "Protecting elections from social media manipulation," *Science*, 2019.
[2] E. Cavaciuti-Wishart, S. Heading, K. Kohler, and S. Zahidi, *The Global Risks Report*. World Economic Forum, 2024.
[3] C. Deane, K. Parker, and J. Gramlich, "Ta year of u.s. public opinion on the coronavirus pandemic," *Pew Research Center*, 2021.
[4] P. S. Hart, S. Chinn, and S. Soroka, "Politicization and polarization in covid-19 news coverage," *Science Communication*, 2020.
[5] R. Sunstein, "The law of group polarization," *UC Law School*, 1999.
[6] H. Maas and L. Dalege, J. Waldorp, "The polarization within and across individuals: the hierarchical Ising opinion model," *J. Compl. Netw.*, 2020.
[7] H. Tajfel and J. Turner, "An integrative theory of intergroup conflict," *The Social Psych. of Intergroup Rel.*, 1979.
[8] P. DiMaggio, J. Evans, and B. Bryson, "Have american's social attitudes become more polarized?" *American journal of Sociology*, 1996.
[9] D. Paschalides, G. Pallis, and M. Dikaiakos, "A framework for the unsupervised modeling and extraction of polarization knowledge from news media," *Trans. Soc. Comput.*, Jan. 2025.
[10] K. Garimella, T. Smith, R. Weiss, and R. West, "Political polarization in online news consumption," *ICWSM*, 2021.
[11] D. Demszky, N. Garg, R. Voigt, J. Zou, J. Shapiro, M. Gentzkow, and D. Jurafsky, "Analyzing polarization in social media: Method and application to tweets on 21 mass shootings," *NAACL*, 2019.
[12] A. Morales, J. Borondo, J. Losada, and R. Benito, "Measuring Political Polarization: Twitter shows the two sides of Venezuela," *Chaos*, 2015.
[13] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, "Quantifying controversy on social media," *Trans. Soc. Comput.*, 2018.
[14] P. Guerra, W. Meira, and C. Cardie, "A measure of polarization on social media networks based on community boundaries," *ICWSM*, 2013.
[15] Z. He, N. Mokhberian, A. Camara, A. Abeliuk, and K. Lerman, "Detecting polarized topics using partisanship-aware contextualized topic embeddings," *EMNLP*, 2021.
[16] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 u.s. election: Divided they blog," *LinkKDD*, 2005.
[17] M. D. Vicario, W. Quattrociocchi, A. Scala, and F. Zollo, "Polarization and fake news: Early warning of potential misinfo. targets," *TWEB*, 2019.
[18] K. Garimella and I. Weber, "A long-term analysis of polarization on twitter," *ICWSM*, 2017.
[19] L. Akoglu, "Quantifying political polarity based on bipartite opinion networks," *ICWSM*, 2014.
[20] R. Balasubramanyan, W. Cohen, D. Pierce, and D. Redlawsk, "Modeling polarizing topics: When do different political communities respond differently to the same news?" *ICWSM*, 2012.
[21] S. Roy and D. Goldwasser, "Weakly supervised learning of nuanced frames for analyzing polarization in news media," in *EMNLP*, 2020.
[22] J. Ørmen and A. Gregersen, "News as narratives," *Oxford Research Encyclopedia of Communication*, 2019.
[23] D. Milne and I. H. Witten, "An effective, low-cost measure of semantic relatedness obtained from wikipedia links," *AAAI*, 2008.
[24] D. Cartwright and F. Harary, "A generalization of heider's theory," *Psychological Review*, 1956.
[25] M. Conover, J. Ratkiewicz, M. Francisco, B. Goncalves, F. Menczer, and A. Flammini, "Political polarization on twitter," *ICWSM*, 2011.
[26] B. Zarei, M. R. Meybodi, and B. Masoumi, "Detecting community structure in signed and unsigned social networks by using weighted label propagation," *Chaos*, 2020.
[27] J. Kerr, C. Panagopoulos, and S. Linden, "Political polarization on covid-19 pandemic response in the united states," *Pers. Individ. Differ.*, 2021.
[28] M. Herrmann and H. Döring, "Party positions from wikipedia classifications of party ideology," *Political Analysis*, 2021.
[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," *NIPS*, 2017.
[30] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?" *Chinese Computational Linguistics*, 2019.