My side, your side and the evidence: Discovering aligned actor groups and the narratives they weave

Pavan Holur[1], **David Chong**[1], **Timothy Tangherlini**[2], and **Vwani Roychowdhury**[1]
[1] Department of Electrical and Computer Engineering, UCLA
[2] Department of Scandinavian, UC Berkeley
{pholur,davidchong13807,vwani}@ucla.edu, tango@berkeley.edu

Abstract

News reports about emerging issues often include several conflicting story lines. Individual stories can be conceptualized as samples from an underlying mixture of competing narratives. The automated identification of these distinct narratives from unstructured text is a fundamental yet difficult task in Computational Linguistics since narratives are often intertwined and only implicitly conveyed in text. In this paper, we consider a more feasible proxy task: Identify the distinct sets of aligned *story actors* responsible for sustaining the issue-specific narratives. Discovering aligned actors, and the groups these alignments create, brings us closer to estimating the narrative that each group represents. With the help of Large Language Models (LLM), we address this task by: (i) Introducing a corpus of text segments rich in narrative content associated with six different current issues; (ii) Introducing a novel two-step graph-based framework that (a) identifies alignments between actors (INCANT) and (b) extracts aligned actor groups using the network structure (TAMPA). Amazon Mechanical Turk evaluations demonstrate the effectiveness of our framework. Across domains, alignment relationships from INCANT are accurate (macro $F1 \geq 0.75$) and actor groups from TAMPA are preferred over 2 non-trivial baseline models ($ACC \geq 0.75$).

1 Background and Motivation

Discussions about current events in public forums involve *consensus building*, with the exchange of beliefs and perspectives producing competing, often conflicting, narratives. A person reading these discussions parses natural language and is able to tease out and maintain representations of the various narratives, including the central actors, their alignments, and the often-contrasting points-of-view presented by the narratives. Replicating this type of comprehension in machines by creating interpretable, mathematical representations of narrative structure is a field of continued computational linguistics efforts (Bailey, 1999; Beatty, 2016). A *narrative* is usually modeled as a *narrative network* of actors (nodes) and their inter-actor relationships (edges). This graph building is, however, a challenging aggregation task since the same narrative can be expressed in natural language in several ways. Conversely, a given text span can include signatures of several underlying narratives.

It is worth noting that a coherent narrative usually features a *small set of critical actors* that emerge through the give and take of online discussions and provides a distilled representation of a particular world view. We refer to these key sets of critical actors that are narratively *aligned* to a shared worldview as *"actor groups"*. People reading or participating in the discussion, in turn, support or even identify with these story actor groups, ensuring the persistence of the narrative in the discussion domain. Identifying these groups of aligned actors is essential to defining the boundaries of a narrative, its current scope and, possibly, its future viability (i.e. if people do not recognize actor groups as central to a narrative, that group and its constitutive members is likely to disappear over time from the narrative space). We therefore consider the detection of actor groups from text as an accessible first step in the larger task of estimating the total narrative structure.

Task: Discovering actor groups from text

Given a corpus of domain-specific free-form text, construct a model to discover the actor groups that undergird the disparate narratives in that domain.

The task of identifying actor groups adds to a growing body of computational linguistics work that identifies salient features of the abstracted narrative structure by exploiting the subtle *contextual clues* available in free-form text: for instance, *In-*

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8938–8952

siders and *Outsiders* (Holur et al., 2022), *Conspiratorial Actors* (Shahsavari et al., 2020b), *Supernodes* and *Contextual Groups* (Tangherlini et al., 2020), and inter-actor event sequencing (Shahsavari et al., 2020a; Holur et al., 2021) (see Related Works Sec. 2 for an extended discussion).

Discovering aligned actors as a means to construct actor groups: A set of mutually aligned actors forms an actor group. Alignment is subtly implied via the inter-actor relationships – often a VERB phrase – in free-range text: Consider, for example, in the news domain of *Gun Regulations* in the United States, a text segment:

{Republicans} $\longrightarrow$ *are funded by* $\longrightarrow$ the {NRA}

suggests {*Republicans, NRA*} are aligned. In contrast, another segment,

{Democrats} $\longrightarrow$ *laid out their anti-* $\longrightarrow$ {Second Amendment} credentials

implies that "Democrats" are *opposed* to the "Second Amendment" and the two actors {*Democrats, Second Amendment*} are disaligned. Tasking a model to discover alignment relationships, a process that comes quite naturally to humans, presents two distinct computational challenges:

Prob-1 Understanding alignment requires human experience: The context traces in language imply but do not explicitly state the alignment between a pair of actors. From the sample text concerning *Gun Regulations*, we observe that the {*NRA* and *Republicans*} were aligned because the NRA *funded* the Republican party; it is *widely accepted* in American politics that *funding signals support*. In another text span, *Democrats* $\rightarrow$ encourage $\rightarrow$ *Black women and men* to vote indicates {Democrats}, {Black women and men} are aligned because *encouragement is a form of validation*. These alignments are trivial to a reader – that offering money and emotional support imply alignment; however, the entire set of phrases that convey alignment in natural language is infinite. Finding the means to map these phrases onto a latent alignment dimension is a fundamental challenge.

Prob-2 Alignment is transitive across a narrative network: Alignment between one pair of actors has the capacity to influence the alignment across other actor pairs in a process that echoes the well-known feature of *Structural Balance Theory* (Cartwright and Harary, 1956; Davis, 1967): a friend of a friend is a friend while an enemy of an enemy is a friend, etc. In the *Gun Regulations* domain, the pair of alignment relationships: "Democrats $\rightarrow$ *sought to ban* $\rightarrow$ the "NRA" (disalignment) and "the Republicans $\rightarrow$ *supported* $\rightarrow$ the NRA" (alignment) *jointly implies* that {Democrats, Republicans} are disaligned, *despite the absence of a direct relationship conveying alignment between them*. Consider a third disalignment: "the NRA $\rightarrow$ *opposed* $\rightarrow$ a gun safety law". Since, {NRA, a gun safety law} are disaligned *and* {NRA, Democrats} are disaligned, according to transitivity, it follows that {Democrats, a gun safety law} *are aligned*. Therefore, modeling this transitivity requires unifying alignment constraints *across* disparate contexts and text spans.

2 Our Approach and Related Work

Computational efforts to address **Prob-1** model human experience by adapting pre-trained large language models (PLM). These models demonstrate considerable semantic *awareness* in several well-known NLP tasks (a product of the knowledge embedded in the exhaustive training corpora); Semantic Role Labeling (SRL) (Zhang et al., 2022), Question-Answering (QA) (Liu et al., 2019a), Sentiment Analysis (SA) (Yin et al., 2020), and Language Generation (LG) (Floridi and Chiriatti, 2020) for instance, all make use of pre-training to boost performance.

The transitivity requirement in **Prob-2** is often addressed by fine-tuning PLMs on biased datasets containing implicit transitivity constraints (Holur et al., 2022; Liu et al., 2019a). Fine-tuning weights encourages generalization *across* data samples. However, these fine-tuned models are dataset-specific and must be retrained for every encountered domain: an expensive and time-intensive task. Alternative approaches use models that are *trained to generate* an external representation of the domain, often in the form of a network (Yu et al., 2022). The network structure enables higher-order consensus insights.

The semantic awareness exhibited by PLMs motivate the adoption of a *transfer learning* approach to extract alignment implicit in text segments. In our work, this is facilitated by Question-Answering (QA) (see Sec. 2.1) that outputs an *alignment network* specific to a conversation domain: the network is a joint representation of individual pairwise alignment relationships between actors (INCANT). Actor groups are identified by exploiting this network structure (TAMPA).

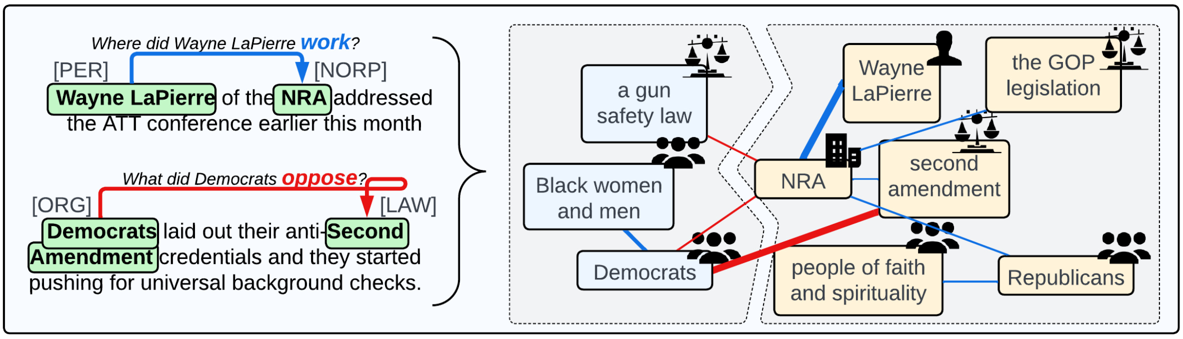


Figure 1: **System overview - Actor subgroups for two opposing narratives identified in the domain of *Gun Regulations*:** Provided a corpus of news articles (left), our approach identifies the likely {actor, question template, answer} alignment relationships that constitute the inter-actor network (right). From here, TAMPA, a custom message-passing algorithm, computes rich actor representations using the alignment score along the edges, and partitions the actors into groups that maintain the individual narratives.

The task of discovering aligned actor groups has strong parallels to identifying *homophily* between users on social media platforms (Khanam et al., 2022). Homophily refers to the tendency for individuals to interact more frequently with those who share similar beliefs and attitudes. Identifying homophilic user groups involves exploiting latent features in the social media with which the users interact; for example, Šćepanović et al. (2017) identifies user cohorts on Twitter by profiling their engagement with political parties; meanwhile Del Tredici et al. (2019) utilizes the neighborhood of a user within a social network to enable inter-user comparison. Our work extends these ethnographic efforts to the narrative landscape: we identify groups of *actors* that *feature in the narrative* using the *contextual alignment clues* present in the language.

2.1 Alignment modeling using question-answering

Recent Natural Language Understanding (NLU) models have implemented a Question-Answering (QA) framework to replicate the iterative process of knowledge acquisition in humans (He et al., 2015; Gutman Music et al., 2022). This framework aims to identify template answer spans that populate a latent knowledge graph, and several network algorithms are applied to infer long-range relationships on this network with multi-hop reasoning and link prediction (Diefenbach et al., 2018; Buck et al., 2017). Similarly, narrative theorists have proposed that questions clarify a narrative's "fillers" or facets (Bailey, 1999). Therefore, a QA-approach to model alignment, an essential facet of narrative structure, should be effective.

Crowdsourcing question templates for alignment retrieval: Deciphering alignment relationships requires asking specialized questions: for example, a reader *knows* that for a *person*-actor (*phrase*), we can identify its alignment constraints by asking: *whom the {phrase} supports*, *whom the {phrase} opposes*, *whom the {phrase} works with*, *what the {phrase} protects/threatens* etc. Typical Question Generation (QG) task setups involve predicting the optimal question given a {*context*} or {*context*,*answer*} tuple (Xiao et al., 2020; Pan et al., 2019); we propose a simple yet effective model to recommend alignment-oriented questions.

Our QG model prioritizes questions conditioned on the NER tag – such as person (PER), or organization (ORG) – of an encountered actor in text. Following an approach similar to He et al. (2015), we crowdsource a basis set of *question templates* (q) and associated alignment score $z_q \in \{-1, -0.25, -0.1, +0.1, +0.25, +1\}$ for each NER tag from $N = 5$ annotators in the en_US locale. $z_q = -1$ indicates that the {*phrase*} actor span in q and the resulting answer are disaligned; a score $z_q = +1$ suggests strong alignment. Popular templates (freq > 2) chosen by annotators are presented in Tab. 1) along with the mode alignment score.

Transfer learning through Question-Answering: We reorient the comprehension abilities of TransformerQA (Liu et al., 2019b), a RoBERTa-large QA PLM trained on the SQuAD dataset (Rajpurkar et al., 2016), to map free-form text relationships to alignment constraints (see **Prob-1**). For an encountered actor s in a text segment x, we identify its NER tag (Honnibal et al., 2020) and associated question templates q_s. *{phrase}* is replaced by s to create a coherent question. Typical QA models

NER Tag	Question Templates
GPE	Who or what was at {phrase}? (0.1), What event took place in or at {phrase}? (0.1)
ORG	Who or what did {phrase} support? (1.0), Who or what did {phrase} oppose? (-1.0), Who worked at {phrase}? (0.25), Where was the {phrase} located? (0.1), Who or what did the {phrase} save? (1.0), Who or what did the{phrase} threaten? (-1.0)
PERSON	Where did {phrase} work? (0.25), Who or what did {phrase} work with? (0.1), Who or what did {phrase} oppose? (-1.0), Where was{phrase} located? (0.1), Who or what did {phrase} support? (1.0)
LAW	What did {phrase} enforce? (1.0), Who or what did {phrase} prosecute? (1.0), Who wrote {phrase}? (1.0), Who or what did {phrase} threaten? (-1.0), Who or what did {phrase} support? (1.0)
NORP	Who or what did {phrase} work with? (0.1), Who or what did {phrase} oppose? (1.0), Who or what did {phrase} support? (1.0)
MONEY	What was {phrase} spent on? (1.0), Where or to whom did the {phrase} go? (0.1), Where or from whom did the {phrase} come from? (0.1)
PRODUCT	What was {phrase} used for? (0.1)
FAC/EVENT	Who or what was at {phrase}? (0.1)
LOC	What event took place at {phrase}? (0.1)

Table 1: **Question templates specific to each NER tag:** Each row contains the frequent question templates (and alignment scores) particular to an entity's NER tag. During runtime, the {phrase} span is replaced by an encountered entity that has a matching NER tag.

learn parameters ϕ such that $p(t|s, q_s; x, \phi)$ is maximized, where t is the correct answer/*answer span* within x. The set of *{subject, question, answer}* tuples form the alignment network (see Section 4.1).

3 Data Collection

News reports are a fertile ground for exploring the formation of opposing narratives and their attendant actor groups since, for any domain, these accounts contain fragments of the various emergent perspectives, the actors aligned with each narrative and the contrast between potentially opposing sides. While individual news articles may favor one narrative perspective over another, a large corpus of articles concerning a single event or issue may, in the aggregate, capture a wide range of these conflicting (sub)narratives. We use a bootstrapped weakly-supervised process to assemble a corpus of such articles particular to a domain C:

1. **Assemble search terms:** A small set of $5-10$ core terms and phrases associated to C_i is manually curated (see Table 2 for domains and seeds);

2. **Discover news articles:** We use GDELT, a Google Jigsaw-powered open real-time news indexing service[1], that pulls recent news articles that match each search term. The search is limited to the en_US locale and to articles published within the last 90 days. GDELT was scraped on Nov 11, 2022. Returned articles are cleaned and common acronyms are resolved[2].

Sample Domain	Seed Terms
Roe v. Wade	roe v wade, abortion, pro-life, pro-choice, clarence thomas, planned parenthood
Gun Regulations	ban on assault weapons, second amendment, gun control, gun rights, mass shooting
War in Ukraine	invasion of ukraine, ukraine-russia war, ukraine, russia, kherson, mccarthy
Vaccine Hesitancy	vaccine hesitancy, vaccine skepticism, vaccine resistance, vaccine refusal, vaccine mandate
Recession Fears	inflation, inflation rate, inflation rates, recession fears, recession, food prices, layoffs, gas prices, OPEC
Immigration	us-mexico border, asylum seekers, immigrants, border wall, visa applications, migrants

Table 2: **Seed terms used for identifying domain-specific news articles:** Each phrase is searched within the GDELT news database and the top articles that match the search terms are scraped using BeautifulSoup (Richardson, 2007) for processing.

We believe there are sufficient actors who are influential in swaying consensus opinion and are unlikely to change their pairwise alignments during the 3-month window. This stabilizes the performance of the inter-actor alignment framework TAMPA (see Sec. 4.2.1) as indicated by our results. The proposed framework also enables identifying actors that switch sides during the observation time window: such actors are positioned by the framework at the outskirts of the core aligned actor groups enabling us to discover the multitude of groups to which they are aligned. For example, we find that in the *Ukraine War* domain, many Republicans were aligned to Russia (Fig. 4). However, Mitt Romney, a moderate Republican, aligns weakly with Ukraine.

The 6 evaluated domains in Tab. 2 have significantly different time frames, ranging from long-standing debates such as *Roe v. Wade* to more recent events like the *War in Ukraine*. These domains also involve a diverse set of actors, range in scope from national issues like *Gun Regulations* to global concerns like *Recession Fears*, and are uni-

[1] www.gdeltproject.org
[2] https://aspe.hhs.gov/common-acronyms

C_i	AC	SC	WC	$\lvert X_i \rvert$	$\lvert V \rvert$	$\lvert E \rvert$
Roe v. Wade	633	15K	363K	6882	4646	5641
Gun Regulation	693	17K	405K	7942	5044	5413
War in Ukraine	774	21K	508K	20940	9248	12769
Vax. Hesitancy	483	13K	337K	6386	4245	4180
Recession Fears	1182	26K	585K	12150	6248	6745
Immigration	690	16K	388K	7706	5933	6923

Table 3: **Data statistics:** AC, SC, WC are the article, sentence, word counts respectively for each evaluated news domain. The number of articles correlate to the popularity of a domain and the seed terms in real-time online GDELT news feeds (from Table 2). Node and edge counts are reported for INCANT networks.

versally recognized as contentious, with multiple viewpoints to consider.

Segmenting the long-form text: Transformer-based models accept a limited token length of context. We split the news articles into smaller segments while retaining many long-range co-reference dependencies:

1. **Auto-regressive coreference resolution:** Each news article is sentence-tokenized $\{s_1, s_2, \ldots, s_N\}$. The auto-regressive seq-2-seq module *greedily* resolves references in a sliding window $k = 5$ $\{s_i, \ldots, s_{i+k}\}$ using a Transformer model trained on OntoNotes 5.0 (Lee et al., 2018). The enriched sentences $\{\hat{s}_i, \hat{s}_2, \ldots, \hat{s}_{i+k}\}$ replace the original set, and the process is repeated after moving the window by a stride $s = 2$. The updated sequence is $\{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_N\}$.

2. **Segment with overlap:** A moving window of length $l = 3$ and stride $d = 2$ partitions $\{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_N\}$ into fragmented shorter sequences to retain sufficient contextual information per segment for inference with downstream Transformer models while remaining computationally feasible at scale.

In this way, we construct X, the set of l-segment spans extracted from news articles specific to domain C. Data statistics for the specific C_is evaluated in this work are presented in Table 3.

4 Methods

4.1 INCANT: The INter-aCtor Alignment NeTwork

We estimate the inter-actor alignment network $G(V, E)$ by identifying the set of relationship tuples R that comprise G. Recall from Section 2.1 that every relationship $r \in R$ is of the form $\{s, q_s, t\}$ The INCANT network estimation process f parameterized by θ estimates the likelihood of each alignment relationship $r := \{s, q_s, t\}$ given a text segment x:

$$p(s, q_s, t|x, \theta) = \underbrace{p(t|s, q_s; x, \theta)}_{C}\underbrace{p(q_s|s; x, \theta)}_{B}\underbrace{p(s|x, \theta)}_{A}. \quad (1)$$

$\{A\}$ $p(s|x, \theta)$: the likelihood of choosing node (actor) s from x. Named Entities (NE) present in x are eligible source nodes and equally likely;

$\{B\}$ $p(q_s|s; x, \theta) := p(q_s|\text{NER}(s); x, \theta)$: the likelihood of choosing a question template q_s from source node s to potential target t: recall that a question template's eligibility is conditioned on the NER tag of s;

$\{C\}$ $p(t|s, q_s; x, \theta)$: the standard Question-Answering (QA) inference task covered in Section 2.1.

For a given text span x, let the set of potential alignment relationships be Φ_x. $|\Phi_x| = NE(x) \times |Q|$ where $NE(\cdot)$ is the set of named entities in x (representing the set of potential source nodes) and Q is the set of all question templates. f_θ assigns a likelihood score to each relationship in Φ_x. Those relationships whose likelihood exceeds a threshold λ $(= 0.7)$ are eligible for constructing the alignment network G_x; the aggregated domain-specific alignment network $G = \cup_{x \in X} G_x$.

G is a signed, multi-edge, directed alignment network. Note that *target* actors, as opposed to *source* actors, need not be named entities. We apply processing steps to G prior to actor group identification: (a) The alignment score z_q corresponding to each question q_s is used as the edge weight (see Tab. 1); (b) Multiple directed edges between a node pair are collapsed into a single undirected edge and the edge weights are averaged; (c) Actors with sparse connectivity (degree $= 1$) are ignored; and (d) The GCC of G is used for further evaluation. Steps (c) and (d) together help to highlight the alignment subnetwork that features the most prominent narratives in the domain. The resulting weighted graph is termed the *INCANT network* particular to domain C and denoted by $\hat{G}$. INCANT network relationships are evaluated using Amazon Mechnical Turk (AMT) (see Tab. 4 for results and Appendix Sec. B for instructions).

4.2 From INCANT networks to actor groups

Given an INCANT network $\hat{G}(\hat{V}, \hat{E})$, we identify the actor groups that constitute the distinct narratives. This task is represented by a partitioning of

$\hat{G}$: we identify $f_{\text{map}} : \hat{V} \to C$, where $c \in C$ is a subset of actors $c \subset \hat{V}$. The narrative subnetwork for c contains those edges $e_{st} \in \hat{E}$ where $\{s, t\} \in c$.

4.2.1 TAMPA: Transitive Alignment via Message PAssing

We describe a framework to construct numerical actor representations that can be compared using a distance measure and clustered. In the context of an INCANT *network* $\hat{G}$, the actor representation learning task translates to learning *node embeddings*. We denote the embedding for node s by $h_s \in \mathbb{R}^D$ ($D = 3$). These embeddings are tuned such that our distance measure, the cosine distance, $d(s,t) = 1 - \frac{h_s \cdot h_t}{||h_s|| ||h_t||}$, $d \in [0, 2]$, is small for aligned actors, and large for opposing ones. Solve:

$$\min_{h_1,\ldots,h_{|\hat{V}|}} \sum_{\{v_1,v_2\} \in \hat{V}} d(v_1, v_2) \times w(v_1, v_2), \quad (2)$$

where w_{v_1,v_2} is the alignment score between v_1, v_2. Solving Eq. 2 directly is intractable since we do not have the alignment constraint between every pair of actors: $\hat{G}$ is data-driven and not fully-connected. To overcome this, we describe a message passing approach to inferring alignment implicit in the network structure. Similar message passing approaches have been shown to be successful in refining node embeddings in the context of co-occurrence networks derived from text (Pujari and Goldwasser, 2021).

Learning graph node embeddings using message passing: The transitive nature of alignment (see **Prob-2**) allows us to define an *effective* alignment score $\tilde{z}_{v_1,v_2}$ between *any* pair of actors $\{v_1, v_2\}$ by considering a *random walk* (Bondy et al., 1976) from v_1 to v_2: for a walk of length L, $\{v_1, t_1, t_2, \ldots, v_2\}$,

$$\tilde{z}_{v_1,v_2} = \gamma^{L-1} \times (z_{v_1} z_{t_1} z_{t_2} \ldots z_{v_2}). \quad (3)$$

γ is a discount factor that takes into account the length of the walk in influencing the alignment between v_1 and v_2. Averaging $\tilde{z}_{v_1,v_2}$ across several random walks provides an estimate of the effective alignment. $\tilde{z}$ approximates w from Eq. 2. Therefore, we can now minimize:

$$\min_{h_1,\ldots,h_{|\hat{V}|}} \sum_{\{v_1,v_2\} \in \hat{V}} d(v_1, v_2) \times \tilde{z}_{v_1,v_2}. \quad (4)$$

Since the loss function in Eq. 4 is non-convex but smooth, we solve for an optimal solution using an iterative method: in every iteration, we sample N random walks per node and the node embedding update is computed using the gradient of the empirical loss. See Appendix Sec. A.1 for details of the parameter gridsearch. Actor groups are generated by clustering TAMPA-trained node embeddings via HDBSCAN (McInnes et al., 2017).

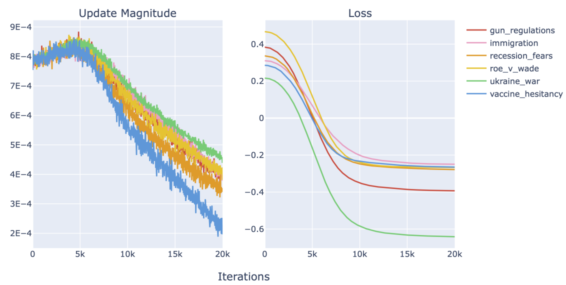


Figure 2: **Convergence plots for TAMPA for different domains:** The update magnitudes reach 0 and the loss function from Eq. 4 converges to an optimal value. See Appendix A.1 for training details.

5 Evaluation and Discussion

To assess the effectiveness of our framework, we use a two-step evaluation approach. First, we rate the quality of the alignment relationships in the INCANT networks. Second, we evaluate the actor groups generated by TAMPA. It is important to note that the inter-actor relationship quality directly impacts the quality of the resulting actor groups. For further reference, we have attached the codebase and supplemental network files. You can access them in our repository at the following link:[3].

5.1 INCANT alignment relationships correlate to human perception

Tab. 4 summarizes the performance of alignment relationship extraction with respect to ground truth labeling performed by MTurk workers (on a random subset of alignment relationships). Details about the labeling setup are provided in Appendix Sec. B. The accuracy, as well as the precision, recall and F1 scores (macro) are high (> 0.75), suggesting good correspondence between the two label sets. Note that in addition to demonstrating that INCANT relationships are accurate, this high performance is indicative of the ability of our QA templates to generalize across domains: The crowd-sourced QA templates in Tab. 1 are not domain-dependent, and yet appear to yield high-fidelity inter-actor alignment relationships for all six evaluated domains.

[3]Repository: `https://osf.io/px3v6`

Domain	ACC↑	P_M↑	R_M↑	$F1_M$↑
Roe v. Wade	0.822	0.797	0.809	0.802
Gun Regulations	0.890	0.872	0.883	0.877
War in Ukraine	0.802	0.793	0.777	0.783
Vaccine Hesitancy	0.843	0.829	0.824	0.826
Recession Fears	0.837	0.778	0.879	0.800
Immigration	0.877	0.860	0.857	0.859

Table 4: **AMT Task 1: Performance of alignment relationship identification:** AMT workers are presented a binarized classification task of identifying whether a pair of actors are aligned or disaligned given input context x. The QA model's predictions are compared to this groundtruth. Workers label a random subset of 1642 samples. The reported precision (P), recall (R), and F1 scores are macro values to account for class imbalance.

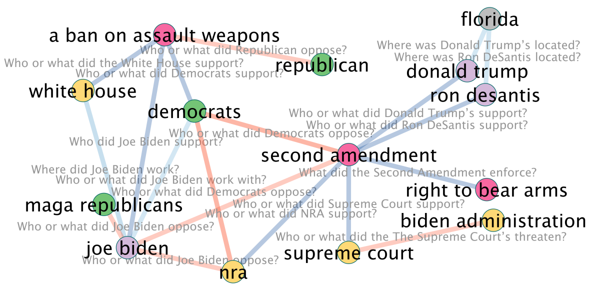


Figure 3: **INCANT subnetwork in the *Gun Regulations* domain:** A subnetwork of 14 high-degree nodes that focuses on a debate about the "second amendment". *Right-leaning* nodes – such as "ron desantis", "donald trump" – support (blue) an unbridled interpretation of the amendment, whereas left-leaning nodes – such as "joe biden" and "democrats" oppose (red) such an interpretation while supporting (blue) a ban on assault weapons.

An INCANT subnetwork for the *Gun Regulations* domain is presented in Fig. 3. An actor's NER tag corresponds to node color, and the question template responsible for an alignment relationship is displayed along the edge. The color intensity of each edge – *blue* (aligned) or *red* (disaligned) – is proportional to the corresponding question template's score (see Tab. 1)[4].

Actor alignments are immediately observed: "donald trump" and "ron desantis" are *aligned* as both actors *support* the "second amendment", and live and campaign in the same state ("florida"). Alignments are *transitive*: {maga republicans, biden} and {biden, the second amendment} are disalignments suggesting maga republicans, second amendment are aligned. TAMPA automates this discovery process.

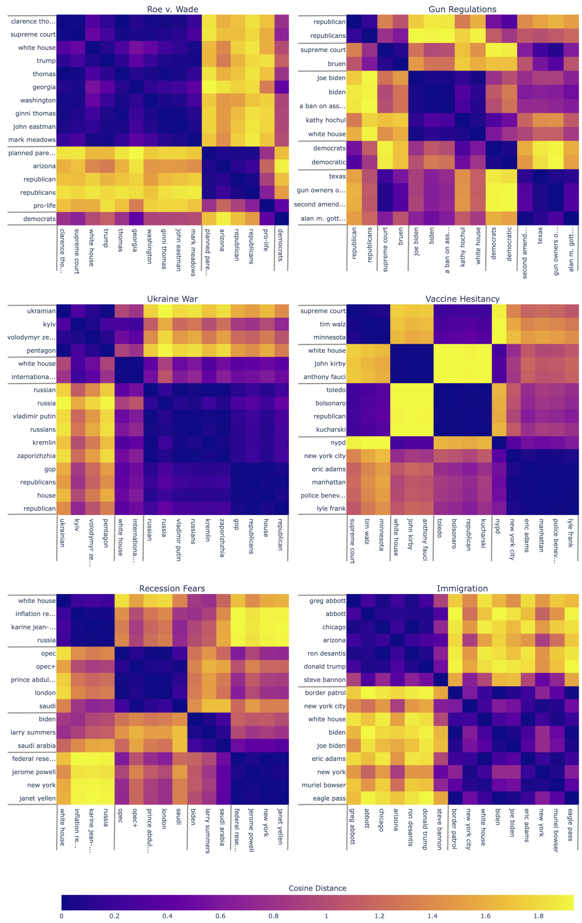


Figure 4: **Actor groups are well-separated in the embedding space:** Each heatmap contains pairwise cosine distances between node embeddings for high-degree actors forming TAMPA actor groups in each evaluated domain. Actors belonging to the same group are listed consecutively (demarcated). The dark diagonal blocks imply that intra-actor group cosine distances are small (purple). In contrast, the inter-actor group distances are large (yellow). One can manually verify that each actor group corresponds to a distinct narrative (see Fig. 5).

5.2 Evaluating actor groups discovered by TAMPA

We first visualize whether the actor groups returned by HDBSCAN are well-separated in the embedding space. As seen in Fig. 4, even the simple measure of pairwise cosine distance is able to separate the clusters. We perform human evaluation (using AMT workers) to evaluate the quality of the actor groups with respect to two baselines.

B1 Community detection: We construct communities in the INCANT network $\hat{G}$ by using the Louvain algorithm (Blondel et al., 2008). Recall that the nodes correspond to actors and each edge corresponds to the question template connecting a pair

[4] All INCANT networks are in the OSF database.

of actors. The alignment scores from Tab. 1 are used as the weights along the edges;

B2 Naïve density-based clustering: Density-based clustering methods such as HDBSCAN can identify clusters provided an adjacency matrix that contains the pairwise distance metric (> 0) between every pair of points. We use the alignment scores for existing edges (similar to **B1**) and replace absent distance values (missing edges in the INCANT network) with a small positive value (2). We replace negative values (between disaligned actors) with a large positive value (10).

In a blind survey, MTurk workers choose the best of three partitionings – **B1**, **B2** and TAMPA – of actors into groups. Labeling setup details are discussed in Appendix Sec. B. Results are presented in Tab. 5: MTurk workers chose TAMPA actor groups to others (baseline $ACC = 0.33$).

Why do we need TAMPA?

Recall that TAMPA was designed to generalize inter-actor alignments beyond the sparse set of direct relationships available in the INCANT networks (sparsity of relationships indicated in Tab. 3). To illustrate TAMPA's operation, we consider the following example involving an actor group identified from the *Recession Fears* domain (narrative description in italics):

{federal reserve, jerome powell, new york, janet yellen}$\rightarrow$ *treasury moves to curb inflation*;

In the corresponding INCANT network $\hat{G}$, there is no direct link between two familiar actors "Janet Yellen" and "Jerome Powell". This is not surprising since our question templates only involve single entities and do not account for multi-entity question, such as "Where have both Janet Yellen and Jerome Powell worked together?". The message-passing scheme in TAMPA addresses precisely this limitation by approximating alignment relationships *transitively*. Consider two alignments that *are* present in the INCANT network:

» "Janet Yellen" (PER) $\rightarrow$ *Who does {phrase} support?* $\rightarrow$ "the Federal Reserve" $\wedge$

» "Jerome Powell" (PER) $\rightarrow$ *Where did {phrase} work?* $\rightarrow$ "the Federal Reserve"

Observe that in these constraints, single-entity QA conveys alignment information with a shared *tertiary* entity: Yellen *supports* the *Federal Reserve* and Jerome Powell *works* at the *Federal Reserve*. TAMPA's message-passing algorithm is incentivized to iteratively refine Yellen and Powell's node representations to be close to that of the Federal Reserve, and effectively construct a strong alignment relationship between the pair:

$\Longrightarrow$ "Janet Yellen", "Jerome Powell" are <u>*aligned*</u>.

Domain	ACC
Roe v. Wade	0.68
Gun Regulations	0.93
War in Ukraine	0.93
Vaccine Hesitancy	0.76
Recession Fears	0.85
Immigration	0.92

Table 5: **AMT Task 2: Performance of TAMPA actor group partitioning vs. B1,B2 baselines:** AMT workers choose the best partitioning of the top 25 actors (by degree) among the 3 models. AMT instructions are presented in the Appendix Sec. B.

Domain	PERM		INCANT		$p_{\text{KS}} \downarrow$
	$\mu \pm$ CI	IQ	$\mu \pm$ CI $\uparrow$	IQ $\downarrow$	
Gun Reg...	0.64 ± 0.03	0.30	0.84 ± 0.02	0.12	8e-31
Immig...	0.66 ± 0.03	0.31	0.80 ± 0.03	0.15	3e-08
Recess...	0.66 ± 0.02	0.31	0.82 ± 0.03	0.14	6e-19
Roe v...	0.59 ± 0.03	0.36	0.81 ± 0.02	0.15	5e-10
Ukraine...	0.60 ± 0.02	0.33	0.72 ± 0.02	0.21	4e-15
Vax. Hes...	0.76 ± 0.03	0.17	0.84 ± 0.03	0.13	4e-05

Table 6: **Silhouette scores – INCANT vs PERM:** μ: The mean Silhouette score across nodes, CI: 95%ile confidence interval, IQ: inter-quartile range, p_{KS}: p-value of the KS-test. These metrics indicate TAMPA creates more distinct clusters in INCANT than PERM.

5.3 Ablation Study: Performance of TAMPA as a function of INCANT network structure

TAMPA relies on the network structure of $\hat{G}$ to model the *effective* alignment between every pair of nodes. We evaluate the extent of this dependence by constructing a modified network baseline, PERM: $\hat{G}$ edges are shuffled while maintaining a constant average node degree. Performance of TAMPA on the INCANT network is compared to the PERM baseline by: (a) Evaluating the separation of actor group clusters using unsupervised metrics; and (b) Visualizing the generated actor groups. As for (b), the random edge shuffling predictably worsens the quality of actor groups since the ground truth alignment information is intentionally corrupted (see Fig. 6 in the Appendix for examples).

For (a), we compute the Silhouette score histogram (Rousseeuw, 1987) $\in [-1, 1]$ after clustering TAMPA-trained node embeddings for both INCANT and PERM: a node's score correlates to its membership strength within its actor group. Strength is computed using the pairwise cosine

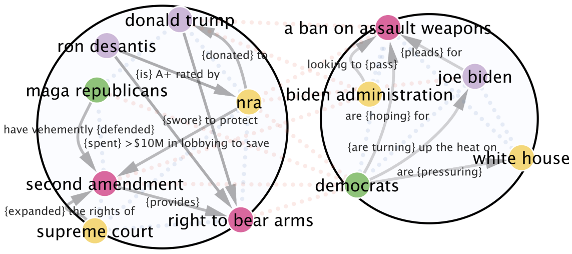


Figure 5: **Distinct narrative networks identified from actor groups in the *Gun Regulations* domain:** The 2 encircled groups are derived by applying TAMPA to the INCANT network in Fig. 3. Directed inter-actor relationships (in grey) are *verb*-phrases identified in news reports between actor pairs. The narrative network on the left promotes the expansion of gun rights; the network on the right seeks to regulate those rights.

distance between the trained embeddings. In Tab. 6, the distribution statistics for the INCANT vs. PERM networks are compared: p_{KS}, the p-value of the Kolmogorov-Smirnov (KS) test (Hodges, 1958) compares the shape of the INCANT vs. PERM histogram distributions. $p < 0.05$ implies the null hypothesis is false, i.e *the two Silhouette score distributions are not identical*. Within each distribution, we compute the mean μ, confidence interval (CI) and the interquartile range (IQ) (Whaley III, 2005): INCANT networks consistently produce a larger μ (close to 1) and smaller IQ, evidence of a score distribution that skews toward 1, and indicative of better separated clusters.

6 Concluding Remarks

In this work, we propose a novel approach for identifying aligned actors and actor groups from the mixture of latent narratives that undergird domain-conditioned free-form text. The success of our approach is evaluated using both qualitative (see Figs. 3,4 and 5) and quantitative (see Tabs. 4,5 and 6) evidence. We show in Fig. 5 that these groups can be used to assemble corresponding narrative networks that convey "my side, your side and the evidence" supporting each side.

When these narrative networks are viewed jointly, we observe a struggle for narrative dominance. In many cases, the tactics proposed in one narrative to counter external threats become threats in and of themselves in other narratives. For instance, in Fig. 5, the relationship tuple "biden administration" $\rightarrow$ *looking to pass* $\rightarrow$ "a ban on assault weapons" (top right) is a *strategy* to counter gun violence, a *threat*. Conversely, this same *strategy* is perceived as a *threat* by gun rights activists. This example highlights the complexity of the narrative landscape and how the same inter-actor relationship can take on distinct, often conflicting roles, depending on the side we choose.

Limitations

Key limitations are listed: (a) Inter-actor networks (from Sec. 4.1) are structured representations of the input data. Since the dataset is assembled on-demand from GDELT, the recall of information given a particular domain depends on its popularity at that time. (b) The TAMPA message passing algorithm (from Sec. 4.2.1) is iterative and converges to a local optimum that may perform poorly with human evaluation for particular domains. (c) The various Transformer models – COREF (from Sec. 3), NER, QA (from Sec. 2.1) – can occasionally produce false positive results. The autoregressive coreference resolution in particular occasionally fails to resolve long-range dependencies across segments, which in turn decreases the recall of nodes and edges from the data. (d) The end-to-end model is only validated for the en_US locale since the Transformer models utilized in the work are most performant in English and many conversation domains are country-specific. (e) In the TAMPA algorithm, actors with a higher degree in $\hat{G}$ are associated to a higher quality of node embeddings since there are more inter-actor alignment constraints. (f) TAMPA uses HDBSCAN as a clustering algorithm: as with any unsupervised ML algorithm, some clusters are more diffuse than others.

Ethics Statement

Process: The datasets used in this analysis were obtained from GDELT, an open-access platform that indexes world news. The scraped dataset is provided in a processed network format, after best-in-class removal of Personally Identifiable Information (PII). Data and codebases are accessible in the OSF repository (`https://osf.io/px3v6/`). *Future Use:* The resulting alignment networks generated by our framework are a representation of the datasets identified on-demand from GDELT. If the sources from GDELT are/or become highly biased to specific news sources, the resulting networks would become biased as well. In this case, the addition of more data sources might be necessary. Additionally, use of this tool in an unmoderated fashion may inhibit free-speech, profile social media users and empower surveillance efforts.

References

Paul Bailey. 1999. Searching for storiness: Story-generation from a reader's perspective. In *Working notes of the Narrative Intelligence Symposium*, pages 157–164.

John Beatty. 2016. What are narratives good for? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 58:33–40.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

John Adrian Bondy, Uppaluri Siva Ramachandra Murty, et al. 1976. *Graph theory with applications*, volume 290. Macmillan London.

Christian Buck, Jannis Bulian, Massimiliano Ciaramita, Wojciech Gajewski, Andrea Gesmundo, Neil Houlsby, and Wei Wang. 2017. Ask the right questions: Active question reformulation with reinforcement learning. *arXiv preprint arXiv:1705.07830*.

Dorwin Cartwright and Frank Harary. 1956. Structural balance: a generalization of heider's theory. *Psychological review*, 63(5):277.

James A. Davis. 1967. Clustering and structural balance in graphs. *Human Relations*, 20(2):181–187.

Marco Del Tredici, Diego Marcheggiani, Sabine Schulte im Walde, and Raquel Fernández. 2019. You shall know a user by the company it keeps: Dynamic representations for social media users in nlp. *arXiv preprint arXiv:1909.00412*.

Dennis Diefenbach, Vanessa Lopez, Kamal Singh, and Pierre Maret. 2018. Core techniques of question answering systems over knowledge bases: a survey. *Knowledge and Information Systems*, 55(3):529–569.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

Maja Gutman Music, Pavan Holur, and Kelly Bulkeley. 2022. Mapping dreams in a computational space: A phrase-level model for analyzing fight/flight and other typical situations in dream reports. *Consciousness and Cognition*, 106:103428.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 643–653.

Joseph L. Hodges. 1958. The significance probability of the smirnov two-sample test. *Arkiv för Matematik*, 3:469–486.

Pavan Holur, Shadi Shahsavari, Ehsan Ebrahimzadeh, Timothy R. Tangherlini, and Vwani Roychowdhury. 2021. Modelling social readers: novel tools for addressing reception from online book reviews. *Royal Society Open Science*, 8(12):210797.

Pavan Holur, Tianyi Wang, Shadi Shahsavari, Timothy Tangherlini, and Vwani Roychowdhury. 2022. Which side are you on? insider-outsider classification in conspiracy-theoretic social media. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4975–4987, Dublin, Ireland. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. *GitHub*.

Kazi Zainab Khanam, Gautam Srivastava, and Vijay Mago. 2022. The homophily principle in social network analysis: A survey. *Multimedia Tools and Applications*.

Kenton Lee, Luheng He, and L. Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *NAACL-HLT*.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. *arXiv preprint arXiv:1905.08949*.

Rajkumar Pujari and Dan Goldwasser. 2021. Understanding politics via contextualized discourse processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1353–1367, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Leonard Richardson. 2007. Beautiful soup documentation. *April*.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Shadi Shahsavari, Ehsan Ebrahimzadeh, Behnam Shahbazi, Misagh Falahi, Pavan Holur, Roja Bandari, Timothy R. Tangherlini, and Vwani Roychowdhury. 2020a. An automated pipeline for character and relationship extraction from readers literary book reviews on goodreads.com. In *12th ACM Conference on Web Science*, WebSci '20, page 277–286, New York, NY, USA. Association for Computing Machinery.

Shadi Shahsavari, Pavan Holur, Tianyi Wang, Timothy R Tangherlini, and Vwani Roychowdhury. 2020b. Conspiracy in the time of corona: automatic detection of emerging covid-19 conspiracy theories in social media and the news. *Journal of computational social science*, 3(2):279–317.

Timothy R Tangherlini, Shadi Shahsavari, Behnam Shahbazi, Ehsan Ebrahimzadeh, and Vwani Roychowdhury. 2020. An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, pizzagate and storytelling on the web. *PloS one*, 15(6):e0233879.

Maxim Tkachenko, Mikhail Malyuk, Nikita Shevchenko, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2021. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Dewey Lonzo Whaley III. 2005. *The interquartile range: Theory and estimation*. Ph.D. thesis, East Tennessee State University.

Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-gen: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation. *arXiv preprint arXiv:2001.11314*.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. SentiBERT: A transferable transformer-based architecture for compositional sentiment semantics. In *Proceedings of the 58th Conference of the Association for Computational Linguistics, ACL 2020, Seattle, USA*.

Donghan Yu, Chenguang Zhu, Yiming Yang, and Michael Zeng. 2022. Jaket: Joint pre-training of knowledge graph and language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11630–11638.

Yu Zhang, Qingrong Xia, Shilin Zhou, Yong Jiang, Guohong Fu, and Min Zhang. 2022. Semantic role labeling as dependency parsing: Exploring latent tree structures inside arguments. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4212–4227, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

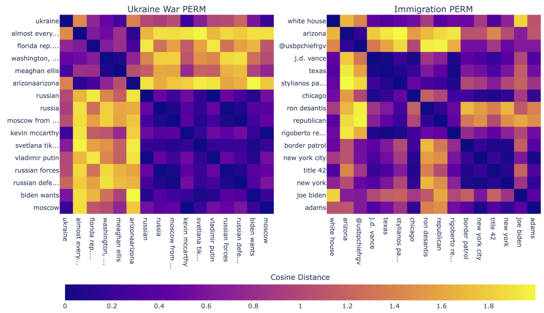


Figure 6: **Pair-wise cosine distances for sample actor groups in the PERM baseline for the domains "Ukraine War" and "Immigration":** Observe that in the permuted baselines, the cosine distances that comprise each block (cluster) have a larger variance (poorly formed clusters) resulting in a lower Silhouette score. Fundamentally, the PERM actor groups are less interpretable in contrast to Fig. 4.

Sanja Šćepanović, Igor Mishkovski, Bruno Gonçalves, Trung Hieu Nguyen, and Pan Hui. 2017. Semantic homophily in online communication: Evidence from twitter. *Online Social Networks and Media*, 2:1–18.

A TAMPA

A.1 Training details

Node embeddings are randomly initialized ($h \in \mathcal{R}^D$, $D = 3$). The length of each random walk N is 10 and $\gamma = 0.95$. Batch size $b = 10$ (the number of random walks considered per node per iteration), number of iterations $M = 20K$. We apply simulated annealing during the learning process: nodes are randomly perturbed with probability $h = 1 - \frac{i}{M}$. Parameter set is presented in Table 7.

Parameter	Values
D	2,**3**,5,10
N	3,**10**,50
b	2,**10**,30,
M	5K,10K,**20K**

Table 7: **Parameter settings for the message passing algorithm:** Optimal choices (by the loss value after convergence) are in bold.

A.2 TAMPA on PERM baseline

See Tab. 7 for the hyperparameters considered for the message passing algorithm TAMPA. The best parameter set is in bold. Models were trained on a 64-core server with 2 TITAN RTX GPUs running Question-Answering and Co-reference Resolution in tandem. Training time per domain does not exceed 1.5 hours.

We are trying to analyze the actors that are *aligned* and *disaligned* in text segments for research purposes. Helping us identify groups of actors that believe in similar worldviews helps us understand how humans perceive influencers and their influence. Your annotations will be used solely for research purposes.

1. Aligned actors share beliefs, have similar motives, have common goals, are supportive of one another, work collaboratively.
2. Disaligned actors oppose each other's beliefs, have counteracting motives, have contrasting goals, are opposed to one another, work competitively (destructively).

We need your help to help us understand alignment in natural language.
Thank you in advance for attempting our survey. We greatly appreciate it!
Make sure to leave this window open as you complete the survey.

Figure 7: **Instructions provided prior to annotator sign-up**: Eligible MTurk workers sign-up for this task after reading this introductory information block describing the annotation task and payment information.

B Instructions to Amazon Mechanical Turk workers

For both Amazon Mechanical Turk (AMT) tasks described below, workers were required to be Masters-granted (https://www.mturk.com/help), present in the en_US locale. Surveys were hosted on-prem and LabelStudio (Tkachenko et al., 2020-2021) was used for creating the survey templates. The post-processed datasets are available for download from the OSF repository (https://osf.io/px3v6/?view_only=b9223fba3e3d4fbcb7ba91da70565604) and are meant for research use with `CC BY 4.0` licensing. Workers were paid $5 for 45 minutes of annotation time. Our estimated time-to-completion was 25 minutes. An overview of the 2 labeling tasks were presented up front to annotators on the AMT platform (see Fig. 7).

B.1 AMT Task 1: Evaluating the quality of alignment relationships

In Fig. 8, we show a snapshot of the instructions presented to MTurk workers to classify a pair of actors present within a context window of text as *aligned* or *disaligned*. Each worker was allowed to label at most 50 samples of the dataset and was allotted 2 hours for the survey. Annotated samples from each worker were randomly sampled and manually verified to ensure quality.

B.2 AMT Task 2: Evaluating the quality of actor groups

MTurk workers are given a preliminary survey to guarantee that they possess sufficient domain knowledge in order to accurately identify the actors that form the actor groups, and to evaluate

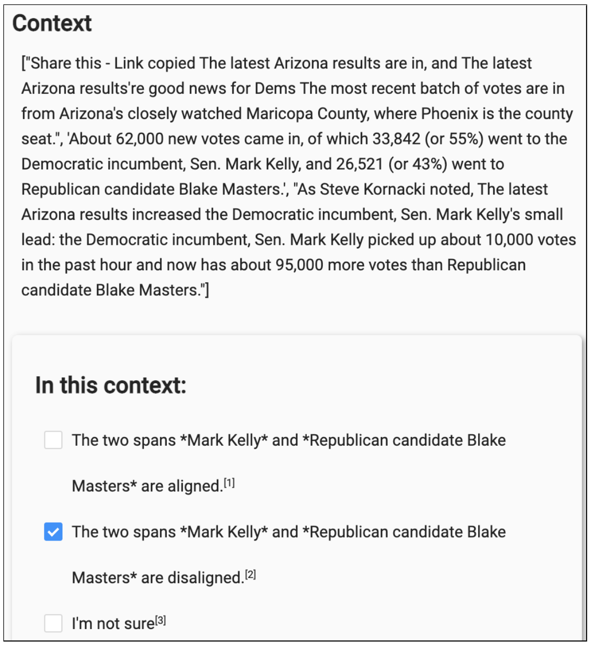
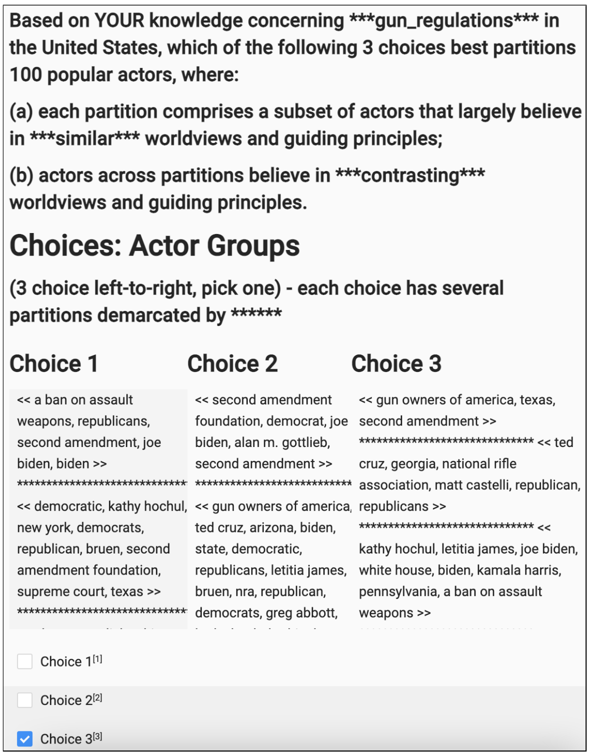


Figure 8: **Labeling instructions for MTurk workers for identifying alignment between actor pairs:** The collected labels are compared to the binarized question template scores $z_s > 0$ in our model. The performance scores are presented in Tab. 4.

whether the actors belonging to each group believe in similar worldviews given the conversation domain. To increase an MTurk worker's chances of being able to identify the actors, we pre-select the top-$K = 25$ actors (by degree) from $\hat{G}$ and their corresponding actor groups. Clustering was performed using the $N = 100$ highest-degree nodes from $\hat{G}$. Fig. 9 shows the instructions provided to the MTurk workers. Once again, annotated samples from each worker were randomly sampled to ensure quality. In total, annotators labeled 240 samples – 40 per domain.



Based on YOUR knowledge concerning ***gun_regulations*** in the United States, which of the following 3 choices best partitions 100 popular actors, where:

(a) each partition comprises a subset of actors that largely believe in ***similar*** worldviews and guiding principles;

(b) actors across partitions believe in ***contrasting*** worldviews and guiding principles.

Choices: Actor Groups

(3 choice left-to-right, pick one) - each choice has several partitions demarcated by ******

Choice 1	Choice 2	Choice 3
<< a ban on assault weapons, republicans, second amendment, joe biden, biden >> ****************************** << democratic, kathy hochul, new york, democrats, republican, bruen, second amendment foundation, supreme court, texas >> ******************************	<< second amendment foundation, democrat, joe biden, alan m. gottlieb, second amendment >> *************************** << gun owners of america, ted cruz, arizona, biden, state, democratic, republicans, letitia james, bruen, nra, republican, democrats, greg abbott,	<< gun owners of america, texas, second amendment >> ****************************** << ted cruz, georgia, national rifle association, matt castelli, republican, republicans >> ****************************** << kathy hochul, letitia james, joe biden, white house, biden, kamala harris, pennsylvania, a ban on assault weapons >>

- [ ] Choice 1[1]
- [ ] Choice 2[2]
- [x] Choice 3[3]

Figure 9: **Labeling instructions for MTurk workers for choosing B1, B2, or TAMPA as the best model for actor group partitioning:** Workers chose one of three partitioning as the best grouping of the top $K = 25$ actors (by degree). Performance scores are presented in Tab. 5. Observe that in this example, the actor groups in Choice 3 – the TAMPA- generated groups – are more semantically coherent than the other options.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
After Concluding Remarks; before References Pg 9. Section number unmarked.*

☑ A2. Did you discuss any potential risks of your work?
After Concluding Remarks; before References Pg 9. Section number unmarked.*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract - Pg 1, Background and Motivation (Introduction) - Pg 1-2

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Sec. 3 - Data Collection (Data), Sec. 4 - Methods (Code)

☑ B1. Did you cite the creators of artifacts you used?
Sec. 3 - Data Collection (Data), Sec. 4 - Methods (Code): Python libraries were used when applicable for data processing and model training. These are referenced in text.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Sec. 3 & See Hyperlink 3.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix Sec. B (Pg. 11)

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Sec. 3

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sec. 3, Appendix Sec. B

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Sec. 3, 4. Tab. 3

C ☑ Did you run computational experiments?

Sec. 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Sec. 4, A.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Sec. 4, A.2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Sec. 6 (6.1, 6.2, 6.3)

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Sec. 3, 4, 5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Sec. 6.1, 6.2

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix Sec. B

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Sec. 6.1, Appendix Sec. B

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Sec. B.1, B.2

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix Sec. B, Sec. 3 (for context)