A Max Flow Min Cut View of Social Media Posts

James Abello[1,2] [a], Timothy R. Tangherlini[3] [b] and Haoyang Zhang[2] [c]

[1] *DIMACS, Rutgers, The State University of New Jersey, 110 Frelinghuysen Rd, Piscataway, New Jersey, U.S.A.*

[2] *Department of Computer Science, Rutgers, The State University of New Jersey, 110 Frelinghuysen Rd, Piscataway, New Jersey, U.S.A.*

[3] *Department of Scandinavian, University of California, Berkeley, 6303 Dwinelle Hall, California, U.S.A.*

Keywords: Social Media, Graph Decompositions, Max Flow Min Cut, Visualization.

Abstract: Viewing social media posts as a collection of directed triples { ⟨ Entity, Verb, Entity ⟩ } provides a frequency labeled graph with vertices comprising the set of entities, and each edge encoding the frequency of co-occurrence of the pair of entities labeled by its linking verb or verb phrase. The set of edges of the underlying topology can be partitioned into maximal subgraphs, called fixed points, each consisting of a sequence of vertex disjoint layers. We exploit this view to observe how information spreads on social media platforms. This is achieved via traces of label propagation across a Max Flow Min Cut decomposition of each fixed point. These traces generate a weighted label set system with an underlined label distribution, from which we derive a barycentric coordinatization of the collection of minimum cuts of each fixed point. This is a novel graph decomposition that incorporates information flow with a multi-layered summary of noisy social media forums, providing a comprehensible yet fine-grained summary of social media conversations.

1 INTRODUCTION

Social media posts and the discussions that they comprise are often noisy and hard to summarize in anything but vague and general terms. A corpus of social media posts such as the Parler data set (Aliapoulios et al., 2021) or Pizzagate (Tangherlini et al., 2020) can be viewed as a collection of directed triplets ⟨ Entity, Verb, Entity ⟩, whose topology is a frequency labeled graph G with vertices comprising the set of entities, and edges encoding the frequency of co-occurrence of pairs of entities labeled by their linking verb or verb phrase. We exploit this view as the initial mechanism to "understand" conversational trends in social posts, and to create multi-scale context-aware summaries of the myriad discussions on a platform. This is achieved via "directed" label propagation between subgraph layers starting with the entities of "local" minimum degree (Abello and Zhang, 2023). Label propagation endows each entity in a particular layer with a set of *cumulative* labels generated by their predecessors in previous layers, which can be interpreted as the incoming local "semantic" content of that entity in the context of the directed "local" subgraph. This incoming "local" semantic content is in turn propagated to its upward neighbors.

Natural questions arise as to how to quantify the textual "semantics" being propagated "upwards" from the lowest layer (i.e. the sources) and how to represent this propagation in a hierarchical manner. Our approach is to compute a binary hierarchy of "Min Cuts" in this local directed subgraph as the mechanism for quantifying the corresponding sequence of textual "Max Flow" label propagations. To achieve this, we assign a capacity to each directed pair of linked entities as a function of their incoming accumulated labels, and compute iteratively the corresponding Max Flow Min Cut decomposition. To summarize the results, we could find a flow weighted longest path in the Max Flow Min Cut hierarchy. We use instead a geometry minimum spanning tree that contains local summaries of the most important flow label sets detected during the Max Flow Min Cut hierarchy computation. We illustrate our method with a small example from the Pizzagate dataset (Tangherlini et al., 2020)(Figure 1).

[a] https://orcid.org/0000-0002-9614-0467
[b] https://orcid.org/0000-0002-1775-2052
[c] https://orcid.org/0000-0002-0082-5876

Abello, J., Tangherlini, T. and Zhang, H.
A Max Flow Min Cut View of Social Media Posts.
DOI: 10.5220/0012862500003756

In *Proceedings of the 13th International Conference on Data Science, Technology and Applications (DATA 2024)*, pages 190-201
ISBN: 978-989-758-707-8; ISSN: 2184-285X

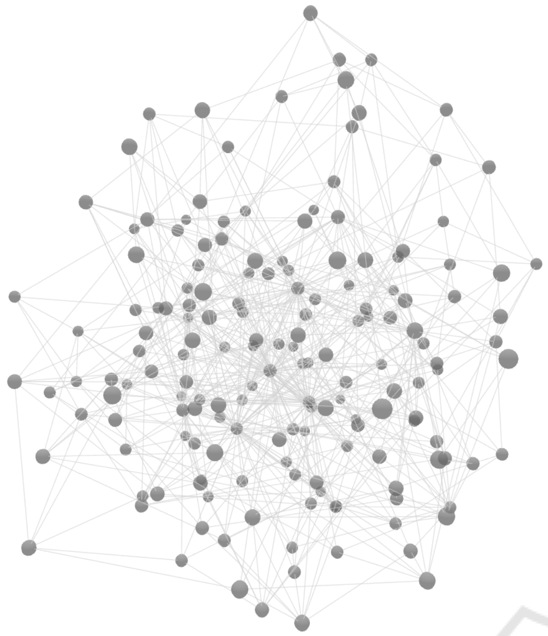

Figure 1: A very small sample input triples topology from the Pizzagate dataset with 166 vertices, 629 edges, and 166 entity labels.

1.1 Paper Contributions

Social information contagion is a phenomenon that can be studied by observing how information spreads on social media platforms. The influence of cognitive biases on polarization has been the subject of extensive research (Hakobyan and Koulovatianos, 2020; Sikder et al., 2020; Peralta et al., 2021; Azzimonti and Fernandes, 2023). One overall goal of this work is to understand regularity at large scales as the collective effects of interactions among pairs of relatively simple entities.

The dynamics of information contagion can be modeled as a process of information spread that is determined by a function of the local neighborhoods of a topology defined by interactions among specially identified "atomic" entities. In the context of social media posts, we propose to view the named entities in the collection of posts as the atomic entities that are linked by verbs. We extract from the posts the collection of ordered triples ⟨ Entity, Verb, Entity ⟩ textually appearing in the posts and we keep counts of their frequency of appearance. In summary, each entity becomes a node in the topology and the directed edges encode pairs of entities in the collection labeled by their interacting verbs and their frequency. Having a well-defined topology allows for the application of graph theoretical methods to ascertain certain properties of the interactions among the entities appearing in the collection.

Our approach consists of several phases.

1. **Data Partition.** Partition the data triplets into maximal subgraphs (called fixed points). The "local" topology of each of these subgraphs is the result of a parameterized iterative neighborhood exploration that starts from a unique "local seed" set (for example, vertices of minimum "local" degree). This local seed set contains the initial "information" being disseminated internally in the chosen subgraph.
2. **Local Label Propagation.** For each subgraph in the partition, to quantify the effect of its seed spreaders on the rest of the "local" topology, we perform one round of directed weighted label propagation, the output of which gets recorded as an information content matrix.
3. **Local Max Flow Min Cut Summaries.** For each subgraph in the partition, a Min Cut decomposition derived from the information content matrix provides a mechanism to quantify and summarize the most influential conversation flows in each subgraph data.
4. **Global Evaluation Summaries.** The aggregation of the local Max Flow Min Cut summaries supports a nuanced, multi-scale evaluation of the underlying conversation space. Although summaries could be generated through the use of an LLM-based chat agent, such as ChatGPT (as we do in the appendix), a human-in-the-loop evaluator is also able to make clear assessments of the conversations, since our graph-based method provides both rich global and multiple local contexts for the conversations, thus providing a "pragmatics" of the discourse arena (Van Dijk, 1980).

The paper layout is as follows. Section 2 details previous work and contains a description of the data sets used. Section 3 describes our computational pipeline that takes a collection of triples coming from social media posts, and provides a barycentric representation of the most representative label flows. Section 4 discusses a sample of our findings in the 2020 US presidential election Parler dataset and the 2017 French presidential election Twitter dataset. It also addresses some limitations of our approach. Section 5 summarizes our conclusions and an appendix contains samples of ChatGPT summaries when prompted with our extracted set of representative labels.

2 PREVIOUS WORK

The study of social media and, in particular, communication in and across social media forums has been a focus of considerable interest, with work on Twitter (Ferrara et al., 2016; Mønsted et al., 2017), Facebook/Meta (Chen et al., 2023; Page et al., 2013), Reddit (Proferes et al., 2021), 4chan/8chan (Baele et al., 2021), and earlier political blogs (Adamic and Glance, 2005) presenting important advances in how to work with this noisy data (in comparison to, for example, mainstream news media) (Tucker et al., 2018). Detecting hate speech and extremism in political blogs has acted as a fundamental goal of much of this work (Aldera et al., 2021). Important in this context is Watts and Dodds' now classic discovery that a large group of easily influenced individuals is more likely to trigger an influence cascade than a small group of well-placed influencers (Watts and Dodds, 2007; Notarmuzi et al., 2022). Consequently, rather than focusing on the pronouncements of a few influencers, it is important to consider the broad discussions on a social forum, and the emerging consensus that guides the discourse arena. The result of many easily influenced people coalescing on a topic of interest is a noisy environment, where multiple strands of conversation are taken up, interrupted, abandoned, reformulated, shouted down, or amplified.

A particular goal of our work is to understand the overall discursive space of a social media forum, predicated on George Boole's famous formulation of what are often referred to in communication studies as "discourse arenas" (Heath and Waymer, 2018): "In every discourse, whether of the mind conversing with its own thoughts, or of the individual in his intercourse with others, there is an assumed or expressed limit within which the subjects of its operation are confined" (Boole, 1854). One of the aims of our work is to estimate the "expressed limit" of these discourse arenas from the data itself, in this case, the posts of users within a forum or a series of forums. Building on the concept of "immanent whole" known from the study of late medieval narratives, where audiences assembled for themselves the episodes and partial stories they heard into a consensus-driven model of the underlying narrative of an epic, saga, or even depictions on a tapestry, we aim to reassemble the underlying narrative framework driving these discussions (Clover, 1986). An understanding of narrative estimation rests on long-standing work in narratology, such as that of Algirdas Greimas (Greimas et al., 1977). The work presented here operationalizes the concept of narrative networks (Bearman and Stovel, 2000), but extends it to provide a multi-scale approach for accessing the semantic richness of the overall space. This approach recognizes that social media often presents stories as partial narratives (Sadler, 2018; Sadler, 2021), and creating entire narratives acts as a form of consensus making. Our approach recognizes the noisy nature of social media, yet also emphasizes the importance of stories in creating communities of belief (Page et al., 2013). Stories are often contingent and dependent on context, and heavily influenced by existing cognitive biases within the community.

The contingent, dynamic nature of social media has been explored in the context of attempts at manipulation of these social spaces. Various tools have been developed to assist in monitoring the impact of, among other things, bots and echo chambers (Davis et al., 2016). These increasingly automated methods provide a complement to more constrained manual narrative extraction methods, even when those are crowdsourced (Kim and Monroy-Hernández, 2015). Magelinski and Carley (Magelinski and Carley, 2022), focusing on Twitter, provide a graph neural network approach that helps define the context-dependent nature of many online conversations while Aiyappa et al. (Aiyappa et al., 2024) reveal how beliefs encoded in these types of conversations can lead to the emergence of both simple and complex contagion dynamics. Work focusing primarily on the content of social media conversations reveals the role that stories and storytelling play on these platforms (Alshaabi et al., 2020). Despite these notable gains on understanding the dynamic processes that drive the emergence of discourse arenas on social media, there is still a pressing need for methods that assist in summarizing, at varying degrees of granularity and with a high degree of context awareness, these noisy conversations. Our work therefore aims to provide a multiscale representation of the dense graphs that derive from the process of estimating the actant interactions in these social media forums; these methods help disentangle the network hairball of social media posts and provide a multi-scalar approach to summarizing the context-based storytelling and could be applied to any domain (Abello et al., 2023).

2.1 Data

The data for this study are derived from two separate datasets, each of which was cleaned and passed through the pipeline described in earlier work (Tangherlini et al., 2020). For illustrative purposes, we also use a dataset describing the Pizzagate conspiracy theory from that work. Our main two datasets focus on the Parler platform in the aftermath of the

2020 US presidential election, and Twitter in the run-up to the 2017 French presidential election. Discussions on Parler focused largely on the emergence of the #Stopthesteal movement and the planning for insurrection on January 6, 2021; the dataset was made available online after the Parler platform was abandoned by service providers (Greenberg, 2021). For the current study, we did not download the user dataset, and only considered the post dataset (Aliapoulios et al., 2021). Usernames were converted to a numerical index, and the mapping between the numerical index and usernames was deleted. Prior to processing, the posts were cleaned for emojis, URLs, images, and two types of text:

1. text that was generated as either welcome messages to the platform, or other platform-related posts, and
2. advertisements.

The posts were sorted and binned by the day that they were posted (normed to UTC).

Posts in these bins were sampled so that, for each day, up to 10% of the day's posts with a minimum of 2000 posts per day were passed to the processing pipeline. Duplicate posts were reduced to a single example (although we kept track of the number of duplicate posts), while posts that included long strings of repetitive text were reduced to the first instance of that text. The resulting corpus consisted of 783,021 posts covering 72 days.

The Twitter dataset for the 2017 French presidential election is the same data as used in (Zhao et al., 2023). A similar binning and cleaning procedure used for the Parler material was used for this data, with the corpus consisting of 1.8M tweets. To facilitate the use of the existing pipeline, we used a neural machine translation step once the dataset was cleaned and binned to convert the French to English with a model tuned for idiomaticity (Smirnov et al., 2022).

The pipeline described in (Tangherlini et al., 2020) creates a series of outputs that constitutes the input for the current work. Foremost among these outputs is a file consisting of triples extracted from the semantic parsing of the posts. Hyperedges representing more complex semantic constructions than simple SVO are decomposed into a series of triples, so that "A gunless takeover by major corporations working with the CCP to bring about a NWO as quickly as possible before they are discovered and remove President Trump from office" is decomposed into a series of nodes and edges: ⟨takeover, working, CCP ⟩; ⟨takeover, bring, NWO ⟩; ⟨corporations, working, CCP ⟩; ⟨ CCP, bring, NWO ⟩; ⟨NWO, remove, Trump ⟩, and so on. This process results in the generation of many relationships for each post. The actants in the sentences, as well as the relationships, are concatenated as described in (Shahsavari et al., 2020), in order to reduce the size of the graph constructed from these extractions.

3 PIPELINE

We create a pipeline of interlocking methods that includes a fixed point edge decomposition, a wave-fragment vertex decomposition, and a Max Flow Min Cut decomposition, as well as a series of methods related to label propagation to generate multi-scale, context-aware summaries of noisy social media discussions (Figure 2). We explain each step of the pipeline below.

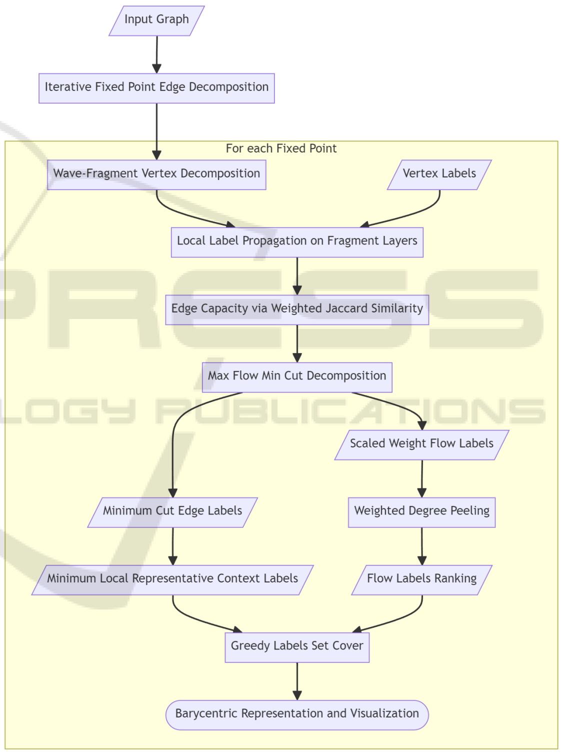

Figure 2: The pipeline of the interlocking methods that take an input graph derived from triples and applies a series of decompositions, including the Max Flow Min Cut algorithm.

3.1 Data Partition

As noted, we view social media posts as a collection of directed triplets ⟨ Entity, Verb, Entity ⟩. The topology of the collection of these posts in a corpus is a frequency labeled graph G with vertices comprising the set of entities extracted from those posts, and edges encoding the frequency of co-occurrence of pairs of entities labeled by the linking verb or verb phrase.

The degree distribution of the underlying graph G can be used to determine a partition of the set of pairs of entities into maximal subgraphs with a local minimum degree, called the peel value of the subgraph, in $O(\sqrt{|E|}|E|)$ time (Abello and Queyroi, 2013). The average degree of each such subgraph is upper bounded by twice its minimum degree. This property provides a layered view of each such subgraph with the lowest layer consisting of the subgraph induced by vertices of minimum degree.

The vertices of each fixed point subgraph can be partitioned into layers with the lowest layer consisting of minimum degree vertices. Special edge maximal subgraphs induced by consecutive layers provide an $O(|E|)$ decomposition of any fixed point into a sequence of "atomic" subgraphs called "fragments" and "waves" (Abello and Nakhimovich, 2022)(Figure 3).

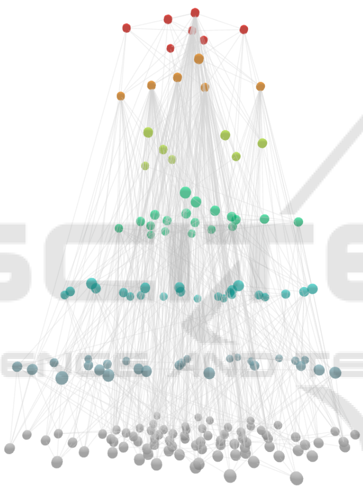

Figure 3: The fragment layer decomposition of the small sample Pizzagate segment in Figure 1 that happens to be a fixed point subgraph of peel value 5.

We exploit this view as the initial mechanism to "understand" conversational trends in social posts. This is achieved via "directed" label propagation between the subgraph layers starting with the entities of "local" minimum degree (Abello and Zhang, 2023).

3.2 Local Label Propagation

We view a fixed point $FP = (V,E)$, as a Directed Acyclic Graph (DAG) $H = (V,E')$ by assigning direction to the edges from lower to higher layers in the fragment decomposition of the fixed point.

We simulate the flow of labels in this DAG by applying the following label propagation schema: Initially, each vertex v is assigned to a multi-set of labels $L(v)$. At each iteration, the multiplicity in the propagated label set $L(v)$ of a vertex is divided by its out-degree and passed to its out-neighbors. The out-neighbors merge the received label sets by summing the multiplicities of each label as their own propagated label set. A pseudocode of the label propagation is shown in Algorithm 1. The time complexity of the label propagation is $O(|l|(|V|+|E'|)) = O(|l||E|)$, where $|l|$ is the number of labels appearing in the fixed point.

Input : DAG $H = (V,E')$, Vertex Label Mapping $L_0 : V \rightarrow 2^{|l|}$
Output: Propagated Label Multi-set Mapping $L : V \rightarrow 2^{|l|}$

$L \leftarrow \{l : 1 | l \in L_0\}$;
for $i \in 1,2,\ldots,k$ **do**
  for $v \in frag_i$ *in fragment ordering* **do**
    $L_{out} \leftarrow L(v)/out-deg(v)$;
    for $u \in$ *out-neighbors of* v **do**
      $L(u) \leftarrow L(u) + L_{out}$;
    end
  end
end
return L;

Algorithm 1: Label Propagation.

Eventually, all sink vertices accumulate the propagated labels. The multi-sets obtained from the label propagation can be viewed as a matrix describing how strongly each vertex v is "influenced" by each label e appearing in all directed paths starting from vertices containing label e. A vertex of high in-degree accumulates more "influence" from its in-neighbors.

3.3 Max Flow Min Cut Decomposition

To quantify the extent to which individual labels contribute to social information contagion, we associate to each directed edge (u,v) with labels $L(u),L(v)$ a capacity value equal to the weighted Jaccard similarity between the multi-sets $L(u),L(v)$. The capacity is thus the ratio of a min-weighted sum of elements in the intersection over a max-weighted sum of elements in the union as in Equation 1.

3.3.1 Labeled Edge Capacity via Weighted Jaccard Similarity

We use the following multi-set version of the Jaccard similarity:

$$WJ(A,B) = \frac{\sum_{e \in A \cap B} \min\{multi(e,A), multi(e,B)\}}{\sum_{e \in A \cup B} \max\{multi(e,A), multi(e,B)\}} \quad (1)$$

where $multi(e,X)$ is the multiplicity of element e in the multi-set X.

$WJ(A,B)$ is ranged from 0 to 1, where 0 means no common element between A and B, and 1 means A and B are identical multi-sets. The weighted Jaccard can be viewed as a weighted sum of all elements in the intersection as $WJ(A,B) = nc(A,B) \sum_{e \in A \cap B} contri(e,A,B)$, where $nc(A,B) = 1/\sum_{e \in A \cup B} \max\{multi(e,A), multi(e,B)\}$ is the normalization constant, and $contri(e,A,B) = \min\{multi(e,A), multi(e,B)\}$ is the contribution of element e to the similarity.

We assign the label capacity $c(u,v)$ of an edge (u,v) in the DAG H as the weighted Jaccard similarity between the propagated label multi-sets $L(u)$ and $L(v)$. The scaled weight of a label e in the intersection of $L(u)$ and $L(v)$ is given by

$$scaled\_weight(e) = nc(L(u), L(v)) contri(e, L(u), L(v)) \tag{2}$$

The time complexity of the weighted Jaccard similarity computation is $O(|l||E|)$.

3.3.2 Iterative Max Flow Min Cut

Given a source and a sink vertex in a directed graph with edge capacities, a Max Flow Min Cut algorithm finds a collection of edges (cut) separating the source and sink vertices with the minimum total capacity, which equals the maximum amount of flow from the source to the sink. Consequently, the directed graph is separated into two subgraphs before and after the cut.

We iteratively apply this algorithm to obtain a hierarchical edge partition of the DAG H. The time complexity of the iterative Max Flow Min Cut decomposition is $O(|frag| maxflow(|V|,|E|))$, where $|frag|$ is the number of wave-fragments in the fixed point, and $maxflow(|V|,|E|)$ is the time complexity of the Max Flow Min Cut algorithm, which is almost linear (Chen et al., 2022). Our real value capacities can be digitized to integers to fit the linear time requirement.

In a cut, the scaled weight of a label e in the intersection before and after the cut is the sum of the scaled weights of e among all the edges in the cut as in Equation 2. The sum of all the labels in the intersection before and after the cut is the total capacity of the cut, which equals the flow of the cut (Figure 4).

Summing over all the cuts, the scaled weight of a label e in the whole fixed point is the sum of the scaled weights of e in all the directed edges in the DAG view of the fixed point. This scaled weight can be viewed as how much information of label e flows in the fixed point.

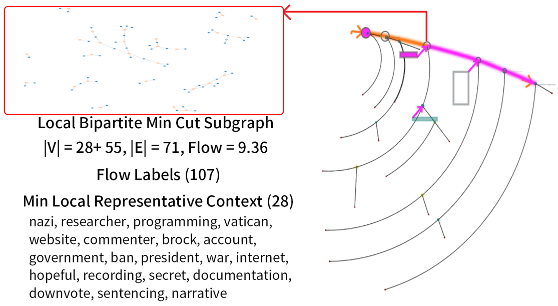

Figure 4: A sample local bipartite Min Cut subgraph in the hierarchy of cuts for the selected fixed point with peel value 5 in Figure 3.

3.4 Summary Extraction

We view the collection of cuts in a fixed point as a weighted set system of their corresponding labels. Each label has a scaled weight indicating its membership in the label set of a cut. We compute a set system weighted peel value for each label appearing in the cut collection. This weighted peel value divided by its weighted degree provides us a mechanism to rank all the labels in non-decreasing order, which describes how much extra information a label contributes to the cuts compared to other labels in the same cuts.

3.4.1 Weighted Set System Peeling

The set system of labels contributing to the cuts in a fixed point can be viewed as a weighted bipartite graph $B = (l \cup C, E_{contri})$, where l is the set of labels and C is the set of cuts, and each edge (e, cut) is weighted by the contribution of the label to the cut. Each vertex v gets weight $w(v)$ equal to the sum of the weights of its adjacent edges. A weighted graph peeling algorithm is applied to B by iteratively removing the vertex with the minimum weighted degree and its incident edges, and updating the weights of its neighbors. This weighted peeling outputs for each vertex its weighted peel value. Labels are ranked non-decreasingly by the ratio of their weighted peel value to their corresponding weighted degree, which describes how much extra information a label contributes to the cuts compared to other labels in the same cuts.

3.4.2 Flow Labels and Representative Label Contexts

For each cut, the *flow labels* of the cut are defined to be the labels that appear both before and after the cut. These labels "glue" the semantics of the contagion before and after the cut. To extract the context of which these flow labels occur, a minimal vertex cover $VC(cut)$ is computed from the local bipartite cut subgraph, and the union of the labels in $VC(cut)$ is taken as the *Minimum Local Representative Context* of the cut. This set of labels "represents" the labels that flow across the cut. To summarize the global contribution of all the labels appearing in the fix point collection, we compute a greedy set cover using the non-decreasing order of the ratio of the weighted peel value over the weighted degree of each label. The ratio is then used as the basis of a barycentric representation. For this representation, the distribution of the ratio of the weighted peel value over the weighted degree of the greedy set cover labels is bucketized by its mean and standard deviation. Each cut is assigned barycentric coordinates according to the contribution of each of its labels to the full fixed point ranked label set. This barycentric representation is subsequently used to create a visualization of the relative influence of the set of labels in a fixed point.

3.4.3 Barycentric Visualization

From a weighted set system $\{S_1, S_2, \ldots S_m\}$ on a finite universe U, one can obtain a ranking of all the elements in U, and an aggregated distribution that is a function of the weight vector of each element $u \in U$ with respect to each of the sets S_i in the set system. This information can be used to create a barycentric visualization. We visualize the representative context labels and the cut collection in a fixed point as a circular barycentric plot. On the circle periphery, the representative context labels are bucketized according to the distribution of the ratio of the weighted peel value over the weighted degree (Figure 5 (a)). Each cut is represented as a point in the interior of the circle with barycentric coordinates proportional to the contribution of each selected label in the cut to the ordered set of buckets on the periphery. A cut with coordinates close to a bucket indicates the bucket contains flow labels that are more representative of the cut labels. A cut will have coordinates that place it closer to the center of the barycentric plot if the cut labels are more broadly spread across the representative context labels. When hovering over a cut, a "lens" that acts as a local magnifying glass onto the cut labels is shown with respect to the same global periphery label order. The interior points of this lens are the local representative context labels of the cut (Figure 5 (b)).

4 DISCUSSION

For a dataset, a minimum spanning tree derived from the barycentric representation of the fixed point de-

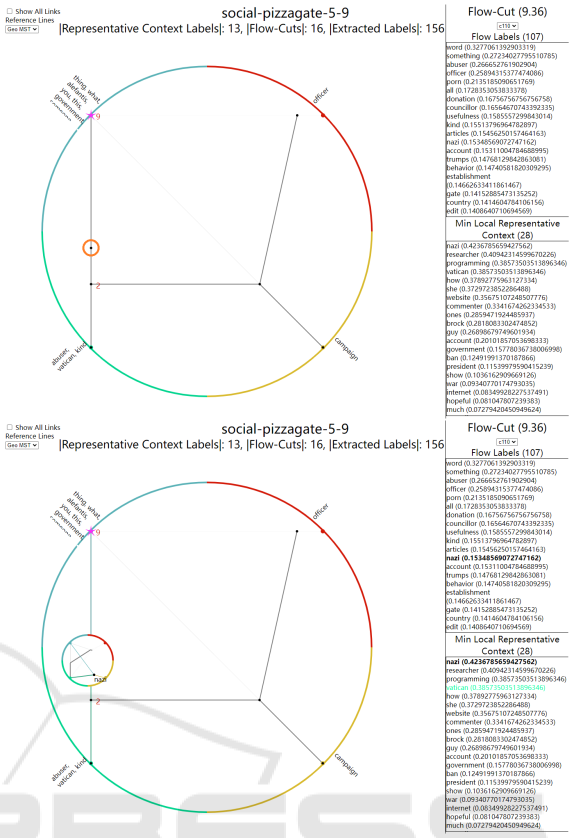

Figure 5: (a) A sample barycentric representation for the Pizzagate segment. The selected cut is circled in orange from Figure 4. (b) A "lens" magnifying the local representative context of the selected cut labels.

compositions captures the important semantic information associated with the nodes on that tree. We create a series of visualizations as described above for each of the fixed points in a dataset to facilitate a macroscopic engagement with the entire corpus space (Börner, 2011; Tangherlini, 2013). In turn, each of these visualizations allows one to engage with the discussions at varying levels of abstraction based on the hierarchical decomposition of the entire narrative framework graph.

Starting from the highest fixed point of the decomposition for the Parler dataset helps illustrate this navigation. Within the circle, we find nodes representing the cuts; along the periphery of the circle are the bucket labels associated with the underlying conversations, effectively laying out the bounds of the discourse arena for this fixed point decomposition. Selecting the MST option allows one to easily navigate this space since once can always find the root, which is denoted with a star. Selecting a node in the visualization not only makes clear the connection between the node and the peripheral labels (each set representing an aspect of the conversation), but also activates

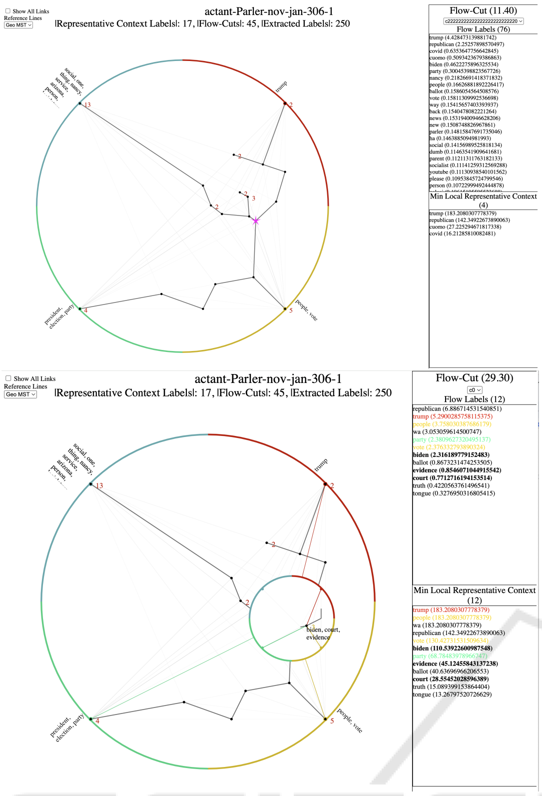

Figure 6: (a) A barycentric visualization of the highest fixed point for the Parler dataset, with the root node marked by a star. The optional MST is selected, connecting the cut nodes to assist navigation. (b) Selecting a node opens a lens onto that cut, and hovering over nodes within that lens highlights words in both the flow labels and the local context labels.

the lens that provides another set of nodes, also in barycentric coordinate space. Hovering over a node inside this lens highlights the related terms in both the flow labels and the local context labels, while presenting the labels on that node (Figure 6).

Exploration of the fixed point decomposition can now proceed in a straightforward manner. Selecting the root of the minimum spanning tree shows a connection to each of the peripheral label groupings, suggesting a uniformity of discussion for this fixed point. Here, the discourse arena is defined by terms (clockwise from top) “Trump”; “people, vote”; “president, election, party”; and “social...nancy...service...Arizona”. The root node, while connected to all periphery groups, is pulled more toward the “Trump” and “people, vote, president” labels, signaling an overall discursive emphasis on Trump and the election itself. Opening the lens onto this node, we find a series of flow labels anchored to the context that raise the specter of “election” (flow) “fraud” (context).

Moving to the right of the root, and thus closer to the same set of peripheral bucket labels (“Trump” and “people, vote”), the discussion is characterized by the flow labels, “republican, Trump, people, Wa[shington], party, vote, Biden, ballot, evidence, court, truth, tongue” and similar context labels, where “tongue” is often found in (a) Biblical quotes and (b) expressions such as “bite your tongue”. Within the lensed view of the node, we find highlighted references to many of the same labels, with one subnode labeled “Biden, court, evidence” which in turn highlights the flow labels “Trump, people, party, vote”, thereby offering an interpretable summary of the local discussion in the context of evidence of a stolen election being used in the courts as a possible intervention, a preoccupation of many users of Parler, and making a clear connection to the root node.

Moving to the left of the root along the MST closer to the left side of the periphery, we find more conversations related to Democrats, Nancy Pelosi and the threat of communism/socialism. Following the minimum spanning tree from the root toward the “Trump” label on the periphery brings up the specter of fraud in the context of the election, with the lensed node “fraud” highlighting the “trump” and “election” context labels, with the addition of “vote” in the flow labels. Further along that path, labels on one of the nodes provides an overview of a discussion of “Pelosi”, the “DOJ”, and “war” in the context of cowardice, suggesting a highly charged reaction to the allegations of fraud, the machinations of Democratic power players and the inaction or possible collusion of the DOJ. Moving around the periphery highlights the links to those cut nodes most associated with the peripheral bucket labels. Interestingly, the node closest to the “people, vote” labels also has a connection to the “social… Nancy… service” labels. One can read this discussion in the context of contrasting ideas of “people” and “country” on one side, with “elite” and its connection to the Democrats and forces of globalization on the other.

The visualizations offer a clear method for developing a series of interlocking overviews of the discussion space. Moving over the nodes and exploring both their local flow labels and the context labels provides an anchoring of these terms. Here, it is not solely that the conversations center around Trump and the election, but rather that these conversations explore allegations of fraud, the role of the Democrats, particularly Kamala Harris, Nancy Pelosi and, not surprisingly, Joe Biden, in helping to orchestrate that fraud, the threats of socialism and communism to America (and Democracy), while also presenting a commentary on the role of “the people” in ensuring that the truth comes out. Social media and the news, includ-

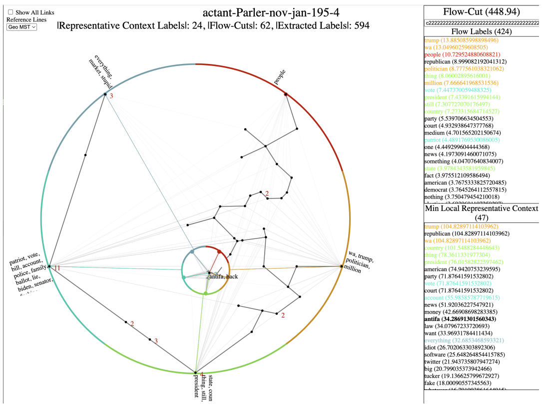

Figure 7: The barycentric visualization of the second highest fixed point for the Parler dataset. The optional MST is selected. Here the discussions are driven by discussions of patriotism, God and threats to Democracy such as Antifa.

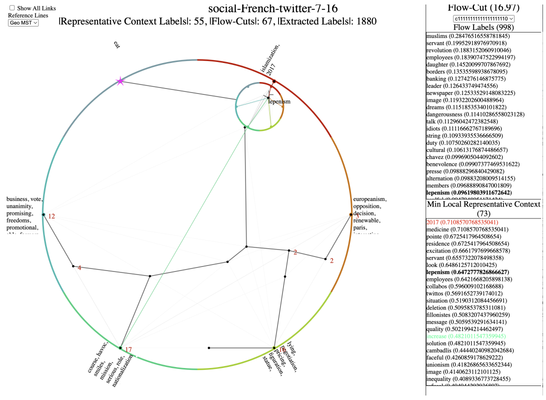

Figure 8: The barycentric visualization of the highest fixed point for the French election dataset. The optional MST is selected. Here the discussions are driven by discussions of Islamization, the EU and Europeanism, nationalization, freedom and business climate.

ing figures such as Tucker Carlson, are clearly in the mix, while the underlying tension about the CoViD-19 pandemic and the role of government actors such as Fauci in the response to the pandemic contribute to the overall conversation space. By interacting with this representation of the fixed point decomposition space, one develops a clearer understanding of the local discussions at each cut, and the local contexts for those discussions. Similarly, one can easily understand the contextually anchored discussions within the broader discourse arena defined by the peripheral bucket labels. Importantly, this process can be repeated for each of the fixed point decompositions.

When one moves to the next fixed point for the Parler dataset, for instance, the conversations become less general, and more focused on aspects of fraud, the power and size of Trump's "patriot" supporters (indicated by intensifiers such as "million"), the role of actants such as "Dominion" and "Democrats", broad groups such as "leftists" and "communists", and the corrupt deep state efforts that characterized the response to the pandemic (Figure 7). A similar navigation process allows one to discover the nature of local discussions and their contexts, while moving left and right along the minimum spanning tree provides a "left-right" context for these conversations, all within the well defined discourse arena of the labeled barycentric coordinate space. The appearance of actant labels from other fixed points helps provide an awareness of the connection between the various levels of more global contexts in which these myriad local discussions are situated.

We applied the same approach to the French presidential election Twitter data described in our data section, and discovered some noteworthy features (Figure 8). The discourse arena is defined by labels of particular salience to the 2017 French presidential election, including concerns about "islamization", the European union, nationalization and the business climate. A node close to the "islamization" label, for instance, provides a discussion with flow labels of "muslims", "servants", "employees", "daughters" and "revolution" in the local context of, among other things, Marine Le Pen's ideology (Le Penism) and labor unions. Another node toward the center of the visualization draws on all the peripheral labels, but has a discussion flow that includes "defeatism", "medicine" and "Moscow" set in the context of "patriotism" and "Le Penism", strongly situating this conversation not only in the context of the challenges Europe confronted in the 2010s but also the rise of right nationalist ideologies such as those of Le Pen. Without knowing much about the French electoral discourse, a navigation of these fixed points allows one to discover not only the main flows of discussion but the various local and global contexts in which they emerged.

4.1 Limitations

Our approach has several limitations. Since the data is noisy, the NLP sentence parsing and semantic role labeling steps can either mis-parse sentences or otherwise fail to find the entities and their relationships. To avoid bottlenecks, we skip failed parses, which tends to reduce the comprehensiveness of the extractions for individual posts. As a result, some of the relationships as well as some of the entities are not as well described as they would be with less noisy and less fragmented posts and discussions. In addition, because we are dealing with triples, complex relationships that cannot be fully decomposed into sets of triples are not captured. The French Twitter dataset raised additional challenges that will be addressed in later iterations of our pipeline. In particular, given current constraints on the relationship extraction section of the pipeline, we discovered that parsing the French sentences natively produced poor results. Since our team had no native French speakers, an important consideration given the idiomatic usages common on Twitter, we relied on a neural machine translation model tuned to French Twitter to provide us with English language equivalents, which were then parsed. This approach unfortunately reduced some of the specificity of those discussions while also garbling others (leading to failed extractions). A native French speaker who was shown the top fixed point of the French election material not only immediately understood the overall discussion space, but also pointed out specific features of the discussions and their contexts, an anecdotal evaluation that provides at least initial support for the methods. Finally, the fixed point decompositions coupled to the max cut min flow algorithms add to the computational complexity of the pipeline, making the code somewhat less transportable than we would like. Given the promise of this approach, we have dedicated future work to a simple implementation of these many working parts.

5 CONCLUSIONS

The threads of conversations on Parler, as in any other discourse arenas, intertwine to form a rich tapestry of situated meaning. But unlike a structured and well-preserved medieval wall-hanging, such as the immediately accessible linear narrative of the Bayeux Tapestry, social media presents as a tangled mess, where linearity and context have fallen prey to speed, volume and, at times incoherence, as if thousands of drunken weavers had sat down to their looms all at once. Disentangling the threads of social forum discussions, and then putting them together in a comprehensible manner that recognizes that meaning is dependent not only on semantics but also the contextual anchoring of those semantics is a non-trivial task. By presenting a hierarchical decomposition of the noisy conversations and their underlying narrative framework graphs, observers may now be able to understand both the local and the global, and all of the levels in between. Ultimately, our method segments conversations and then links them back together in a more coherent order derived entirely by the underlying data itself. In short, it is not particularly useful if all one can conclude is that Parler was focused on discussions about an alleged stolen election. Instead, it is imperative to have methods for finding the local debates within the broader corpus, the discussions about threats and threateners, and the emerging consensus on a strategy to #stopthesteal involving violent insurrection. It is, afterall, one thing to say that the Bayeux Tapestry is about events leading to the Norman Conquests of 1066. It is another thing entirely to be able to see each of the small scenes on the tapestry in all of their various contexts and understand the possible political drivers of that particular version of events.

ACKNOWLEDGEMENTS

We would like to thank Peter Broadwell for his many contributions developing earlier versions of this work. The Parler dataset was initially cleaned and prepared by Alexandre Miller, Megan Leng, Olivia Kris and Chandini Dialani.

REFERENCES

Abello, J., Broadwell, P. M., Tangherlini, T. R., and Zhang, H. (2023). Disentangling the folklore hairball: A network approach to the characterization of a large folktale corpus. *Fabula*, 64(1-2):64–91.

Abello, J. and Nakhimovich, D. (2022). Graph waves. *Big Data Res.*, 29:100327.

Abello, J. and Queyroi, F. (2013). Fixed points of graph peeling. *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 256–263.

Abello, J. and Zhang, H. (2023). Graph peeling semantics. In *EDBT/ICDT Workshops*.

Adamic, L. A. and Glance, N. S. (2005). The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD '05*.

Aiyappa, R., Flammini, A., and Ahn, Y.-Y. (2024). Emergence of simple and complex contagion dynamics from weighted belief networks. *Science Advances*, 10.

Aldera, S., Emam, A., Al-Qurishi, M., Alrubaian, M., and Alothaim, A. (2021). Online extremism detection in textual content: A systematic literature review. *IEEE Access*, 9:42384–42396.

Aliapoulios, M. M., Bevensee, E., Blackburn, J., Bradlyn, B., Cristofaro, E. D., Stringhini, G., and Zannettou, S. (2021). A large open dataset from the parler social network. In *International Conference on Web and Social Media*.

Alshaabi, T., Adams, J. L., Arnold, M. V., Minot, J. R., Dewhurst, D. R., Reagan, A. J., Danforth, C. M., and Dodds, P. S. (2020). Storywrangler: A massive exploratorium for sociolinguistic, cultural, socioeconomic, and political timelines using twitter. *Science Advances*, 7.

Azzimonti, M. and Fernandes, M. (2023). Social media networks, fake news, and polarization. *European journal of political economy*, 76:102256.

Baele, S. J., Brace, L., and Coan, T. G. (2021). Variations on a theme? comparing 4chan, 8kun, and other chans' far-right "/pol" boards. *Perspectives on Terrorism*, 15(1):65–80.

Bearman, P. S. and Stovel, K. (2000). Becoming a nazi: A model for narrative networks. *Poetics*, 27:69–90.

Boole, G. (1854). *The Laws of Thought (1854)*. The Open court publishing company, London,.

Börner, K. (2011). Plug-and-play macroscopes. *Communications of the ACM*, 54(3):60–69.

Chen, B., Lukito, J., and Koo, G. H. (2023). Comparing the #stopthesteal movement across multiple platforms: Differentiating discourse on facebook, twitter, and parler. *Social Media + Society*.

Chen, L., Kyng, R., Liu, Y. P., Peng, R., Gutenberg, M. P., and Sachdeva, S. (2022). Maximum flow and minimum-cost flow in almost-linear time. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 612–623.

Clover, C. (1986). The long prose form. *Arkiv för Nordisk Filologi*, 101:11–39.

Davis, C. A., Ciampaglia, G. L., Aiello, L. M., Chung, K., Conover, M. D., Ferrara, E., Flammini, A., Fox, G., Gao, X., Gonçalves, B., Grabowicz, P. A., Hong, K., Hui, P.-M., McCaulay, S., McKelvey, K. R., Meiss, M. R., Patil, S., Kankanamalage, C. P., Pentchev, V., Qiu, J., Ratkiewicz, J., Rudnick, A., Serrette, B., Shiralkar, P., Varol, O., Weng, L., Wu, T.-L., Younge, A. J., and Menczer, F. (2016). Osome: the iuni observatory on social media. *PeerJ Prepr.*, 4:e2008.

Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The rise of social bots. *Commun. ACM*, 59(7):96–104.

Greenberg, A. (2021). An absurdly basic bug let anyone grab all of parler's data. *Wired (12 January 2021)*.

Greimas, Porter, and Cathérine (1977). Elements of a narrative grammar. *Diacritics*, 7:23.

Hakobyan, Z. and Koulovatianos, C. (2020). Populism and polarization in social media without fake news: The vicious circle of biases, beliefs and network homophily. *Higher School of Economics Research Paper No. WP BRP*, 227.

Heath, R. L. and Waymer, D. (2018). Terministic dialectics of individual and community agency: Co-creating and co-enacting public interest. In *Public Interest Communication*, pages 32–51. Routledge.

Kim, J. and Monroy-Hernández, A. (2015). Storia: Summarizing social media content based on narrative theory using crowdsourcing. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*.

Magelinski, T. and Carley, K. M. (2022). Contextualizing online conversational networks. *ArXiv*, abs/2207.13055.

Mønsted, B. M., Sapiezynski, P., Ferrara, E., and Lehmann, S. (2017). Evidence of complex contagion of information in social media: An experiment using twitter bots. *PLoS ONE*, 12.

Notarmuzi, D., Castellano, C., Flammini, A., Mazzilli, D., and Radicchi, F. (2022). Universality, criticality and complexity of information propagation in social media. *Nature communications*, 13(1):1308.

Page, R. E., Harper, R., and Frobenius, M. (2013). From small stories to networked narrative: The evolution of personal narratives in facebook status updates. *Narrative Inquiry*, 23:192–213.

Peralta, A. F., Neri, M., Kertész, J., and Iñiguez, G. (2021). Effect of algorithmic bias and network structure on coexistence, consensus, and polarization of opinions. *Physical Review E*, 104(4):044312.

Proferes, N., Jones, N., Gilbert, S. A., Fiesler, C., and Zimmer, M. (2021). Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society*, 7.

Sadler, N. (2018). Narrative and interpretation on twitter: Reading tweets by telling stories. *New Media & Society*, 20:3266 – 3282.

Sadler, N. (2021). *Fragmented narrative: Telling and interpreting stories in the Twitter age*. Routledge.

Shahsavari, S., Holur, P., Wang, T., Tangherlini, T. R., and Roychowdhury, V. P. (2020). 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*, 3:279 – 317.

Sikder, O., Smith, R. E., Vivo, P., and Livan, G. (2020). A minimalistic model of bias, polarization and misinformation in social networks. *Scientific reports*, 10(1):5493.

Smirnov, A. V., Teslya, N., Shilov, N., Frank, D., Minina, E., and Kovacs, M. (2022). Comparative analysis of neural translation models based on transformers architecture. In *International Conference on Enterprise Information Systems*.

Tangherlini, T. R. (2013). The folklore macroscope: Challenges for a computational folkloristics. *Western Folklore*, pages 7–27.

Tangherlini, T. R., Shahsavari, S., Shahbazi, B., Ebrahimzadeh, E., and Roychowdhury, V. P. (2020). An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, pizzagate and storytelling on the web. *PLoS ONE*, 15.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., and Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.

Van Dijk, T. A. (1980). The semantics and pragmatics of functional coherence in discourse. *Speech act theory: Ten years later*, pages 49–65.

Watts, D. J. and Dodds, P. S. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34:441–458.

Zhao, W., Guo, F., Lerman, K., and Ahn, Y.-Y. (2023). Discovering collective narratives shifts in online discussions. *ArXiv*, abs/2307.08541.

APPENDIX

As an experimental approach to assist with broader summarization, we presented an LLM-based chat agent (here, ChatGPT) with a series of bucket labels, cut labels and flow labels. For the first prompt, using the highest fixed point on the Parler data, we passed the agent the following prompt, limiting the response to 250 words:

> *Create a brief summary of the discourse arena represented by the following label buckets:* trump; people, vote; president, election, party; social, one, thing, service, nancy, arizona, propaganda.

The response took the form of a useful summary of this discourse arena:

> The discourse arena outlined by your labels focuses on U.S. political and social themes. The "Trump" bucket suggests discussions related to former President Donald Trump, including his political influence and policies. The "people, vote" bucket centers on voter engagement and the democratic process, emphasizing public participation in elections. The "president, election, party" bucket likely involves debates around presidential elections, political party dynamics, and electoral strategies. Finally, the "social, one, thing, service, nancy, arizona, propaganda" bucket covers broader social issues, possibly including social services, localized concerns in Arizona, and media influence, with "Nancy" potentially referring to a public figure like Nancy Pelosi. This arena melds discussions on political figures, voter behavior, and media's role in politics.

We then passed it the top context labels on the root of the minimum spanning tree, with the following prompt:

> *In one of the related discussions, I find the following ranked labels*: trump, people, wa[shington], republican, vote, biden, party, evidence, ballot, court, truth. *Provide a contextualized summary of no more than 100 words as to what they are talking about in the context of the discourse arena you characterized above.*

We received the following response:

> In this discussion, the focus remains on political dynamics within the characterized discourse arena, with prominent mentions of Trump, the Republican Party, and the electoral process. The inclusion of "evidence," "ballot," and "court" suggests ongoing debates over electoral integrity and legal challenges, reflecting concerns about the legitimacy of election outcomes. "Truth" likely denotes discussions surrounding conflicting narratives or interpretations of events. Overall, the conversation revolves around the intersection of politics, democracy, and media influence, echoing broader themes of political polarization and scrutiny over electoral processes.

We then chose another cut with its context labels, and passed it a similar prompt:

> *Another localized discussion has the labels*: trump, election, american, fraud, fucking, tucker, lady, oh, official, freedom, free, speech, butt, yes, heart, kamala, usa, thousand. *How does that resolve in the context of both the discourse arena and the conversation you saw earlier? Again, give me your thoughts in 100 words or less.*

And we received the following response:

> This localized discussion appears more charged and confrontational within the broader political discourse. Central to the conversation are claims of electoral fraud linked with Trump, reflecting a contentious debate over the integrity of elections. The inclusion of "Tucker" might suggest commentary or influence from media figures like Tucker Carlson, enhancing the discussion's polarized nature. Words like "fucking," "freedom," "free speech," and the mention of other figures like "Kamala" indicate a highly emotive and possibly divisive exchange concerning American democratic values and freedoms, aligning with the earlier noted media influence and political rivalries in the discourse arena.

A Closing Thought:
Although this approach is not systematic in its current form, it could be easily addressed programmatically. The usefulness of these summaries is notable and, in conjunction with our work, provide the basis for an easily interpretable, macroscopic engagement with large, complex and noisy social media platforms. Creating a method combining an LLM chat agent with a RAG (retrieval augmented generation), for instance, could potentially further tune the generated responses to the overall space of the conversations. We reserve these refinements for future work.