How Ideas Flow across Multiple Social Groups

Xiting Wang* Shixia Liu* Yang Chen* Tai-Quan Peng† Jing Su‡ Jing Yang§ Baining Guo¶

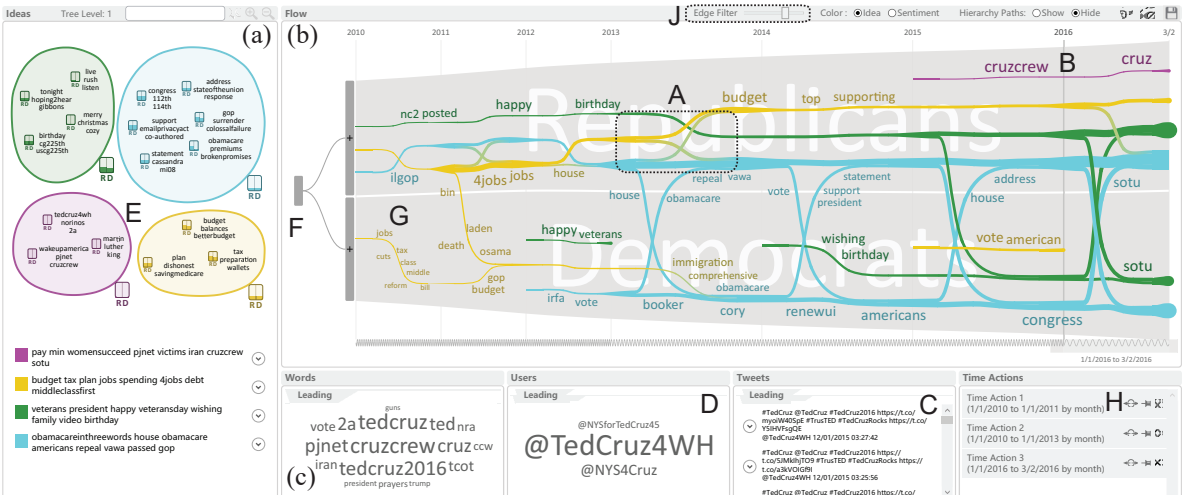

Figure 1: Top-level idea clusters and the corresponding flows of the Congress data: (a) an overview of ideas; (b) lead-lag relationships over time; (c) information panel.

Abstract

Tracking how correlated ideas flow within and across multiple social groups facilitates the understanding of the transfer of information, opinions, and thoughts on social media. In this paper, we present IdeaFlow, a visual analytics system for analyzing the lead-lag changes within and across pre-defined social groups regarding a specific set of correlated ideas, each of which is described by a set of words. To model idea flows accurately, we develop a random-walk-based correlation model and integrate it with Bayesian conditional cointegration and a tensor-based technique. To convey complex lead-lag relationships over time, IdeaFlow combines the strengths of a bubble tree, a flow map, and a timeline. In particular, we develop a Voronoi-treemap-based bubble tree to help users get an overview of a set of ideas quickly. A correlated-clustering-based layout algorithm is used to simultaneously generate multiple flow maps with less ambiguity. We also introduce a focus+context timeline to explore huge amounts of temporal data at different levels of time granularity. Quantitative evaluation and case studies demonstrate the accuracy and effectiveness of IdeaFlow.

Keywords: Idea flow, lead-lag, focus+context, correlated clustering, flow map.

*X. Wang, S. Liu, and Y. Chen are with School of Software, Tsinghua University. S. Liu is the corresponding author. E-mail: {thu.xt.wang, chen1984yang}@gmail.com, shixia@tsinghua.edu.cn.

†T.-Q. Peng is with Michigan State University. E-mail: winsonpeng @gmail.com.

‡J. Su is with Tsinghua University. E-mail: sujingxw@tsinghua.edu.cn.

§J. Yang is with UNCC. E-mail: jing.yang@uncc.edu.

¶B. Guo is with Microsoft Research. E-mail: bainguo@microsoft.com.

1 Introduction

According to Webster's dictionary, an idea is "a formulated thought or opinion." Currently, we are faced with a highly complex information environment as social media has engaged the general public in the production and exchange process of various ideas. These ideas, each of which can be characterized by a set of words, have demonstrated different life-cycle patterns on social media. Some ideas are very enduring and energetic, and can flow within and across multiple social groups over a long period of time, while other ideas are quite ephemeral and stagnant, and can only stay within a specific group of users for a limited period of time. Tracking the way ideas flow on social media is of theoretical and practical significance for social scientists, policy-makers, and business managers, because it can enrich our knowledge about the emergence and evolution of a social idea, advance our understanding of the convergence and divergence of multiple social ideas, and sharpen our insights into the roles of various social groups in the idea flow process.

For example, public health researchers often analyze emerging health threats like Ebola and aim to develop an efficient and effective action plan to tackle such threats. By analyzing the lead-lag relationships between a set of correlated ideas mined from the Ebola dataset, they can identify the most influential social groups and make suggestions to the government. The lead-lag relationships refer to instances where an idea discussed by individuals in group A (lead) is followed by those in group B (lags). Correlation is a statistical technique that shows whether and how strongly pairs of variables are related [52]. In our work, two ideas are correlated if they frequently propagate information concerning each other (measured by meta-data such as re-tweets), have similar content, and/or exhibit similar temporal changes in terms of given quantitative values such as word frequency. Compared with topics, ideas are more correlated with each other. Another example is to disclose the leadership and impact of different groups of individuals on multiple social issues. The idea flows among such individuals can reveal the lead and lag relationships between political parties on a set of correlated ideas. Better understanding

such flow information can facilitate both parties' timely responses to rival voices and help shape their political campaigns. Driven by these applications, there is an increasing demand in scholarly and practical research to track such idea flows, especially the lead-lag relationships, within and across groups.

Tracking idea flows from the huge amount of data collected on social media is technically demanding. There are two major challenges that we need to address. One is to model ideas and track their lead-lag relationships over time. Existing methods either employ word-based tweet content [21] or cointegration between word time series [53] to model idea flows. Generally, the model accuracy is unsatisfactory. Moreover, in real-world applications, it is desirable to track the lead-lag changes across multiple social groups regarding a specific set of correlated ideas. Due to the lack of a consistent method to jointly estimate lead-lag relationships from multiple social groups, the existing methods either track correlated idea flows between two groups [53] or model lead-lag relationships among multiple groups on a specific idea instead of a set of correlated ideas [21]. The second challenge is to design an intuitive visualization to illustrate idea flows across a long time span. The idea flow model often produces hundreds of ideas and thousands of lead-lag relationships. Trying to display all of them at once will cause visual clutter and ambiguity. Furthermore, a dataset may contain thousands of time points. As such, an easy-to-understand metaphor is necessary to efficiently and effectively represent the idea flow over a long timeframe, which allows experts to focus on a selected time period of interest while keeping others in context.

We have developed a visual analytics system, IdeaFlow, to tackle these challenges. To model idea flows accurately, we have developed a random-walk-based correlation model and integrated it with Bayesian conditional cointegration [5] and a tensor-based technique [18]. Specifically, we first calculate correlations between word time series by using Bayesian conditional cointegration. The correlations between word time series are then fed into the random-walk-based correlation model, which combines correlations between word time series, the tweet content, and meta-data (e.g., re-tweets) to generate more accurate temporal correlations of words. Next, we use tensor decomposition to cluster the words into ideas and aggregate the temporal correlations between words into lead-lag relationships. To effectively convey complex lead-lag relationships, we have developed a visualization that combines the strengths of a Voronoi-treemap-based bubble tree, a flow map, and a timeline. In particular, we have developed a Voronoi-treemap-based bubble tree to help users quickly get an overview of a set of ideas. A correlated-clustering-based layout algorithm is used to simultaneously generate multiple flow maps with less ambiguity. We have also introduced a focus+context timeline to explore huge amounts of temporal data at different levels of time granularity. Based on the three visualizations, IdeaFlow offers multiple coordinated visualizations facilitated by interactive explorations.

The major contributions of this work are:

- **A visual analytics system** that helps experts understand and analyze the lead-lag changes within and across social groups regarding a specific set of correlated ideas.
- **A correlation model** that substantially improves model accuracy by taking into account the tweet content, meta-data, and the cointegration between word time series.
- **A coordinated visualization** that combines the strengths of a correlated-clustering-based flow map, a focus+context timeline, and a Voronoi-treemap-based bubble tree to convey lead-lag at different levels of granularity and with less ambiguity.

2 Related Work

2.1 Mining Lead-lag Relationships

Our idea flow tracking algorithm is related to lead-lag analysis. Current literature on lead-lag analysis can be classified into two categories: global lead-lag analysis that models lead-lag relationships across all time points and local lead-lag analysis that models lead-lag relationships at each time point. Global lead-lag analysis methods [19, 48] have achieved a certain amount of success in identifying leading (or most influential) corpora and documents. However, they are not able to detect local lead-lag changes over time.

Recently, some local lead-lag analysis methods have been developed. For example, TextPioneer [21] learns local lead-lag relationships among multiple corpora by calculating the content correlation at different time points. More specifically, corpus **A** is considered to be the local lead at time point t if its content at t is more similar to the future rather than the past content of corpus **B**. However, this method only detects the lead-lag changes of the same topic. Zhong et al. [53] modeled local lead-lag relationships of a set of correlated ideas by detecting the cointegration between word time series. Specifically, they represented each word by a time series that encodes its frequency change over time. Two word time series are considered to be cointegrated (or correlated) in a time period if they have a stationary linear combination during that period. However, this method limits the learning of the lead-lag relationships to only two groups. Compared with this method, our model is able to detect local lead-lag changes within or across multiple groups on a set of correlated ideas. Our method substantially improves the model accuracy by developing a random-walk-based correlation model in which the content, meta-data (e.g., follower-followee relationships), and correlations between word time series are taken into account. Furthermore, we have developed a visual analytics system that empowers users to better analyze and understand lead-lag changes within and across multiple groups.

2.2 Visual Analysis of Topic Influence

The problem of visually tracking and analyzing topic evolution in large text corpora has received considerable attention in recent years [22, 27, 38]. For example, TIARA [26, 35] and Visual Backchannel [11] visually depict how topics or text clusters evolve over time using a river metaphor. HierarchicalTopics [14] extracts hierarchical structures of topics using Bayesian rose trees [23, 37] and visually tracks their temporal changes in a hierarchical ThemeRiver visualization. ParallelTopics [12] integrates an LDA topic-modeling algorithm with a parallel coordinate metaphor to represent document topic distributions and their temporal evolution.

The evolving topics may be correlated with other topics over time by a variety of relationships. The most interesting relationships are topic popularity [13, 28], topic competition and collation [39, 50], and topic merging and splitting [8, 9, 15, 25]. For example, EventRiver [28] and Leadline [13] visually depict the temporal popularity of document clusters (known as events) in a text corpus. Some recent research focuses on visually analyzing the competition relationships among topics [50], as well as the coopetition [39] of topics as their influence on public attention. They were visually conveyed in time-based flow visualizations. TextFlow [8] was developed to visually convey topic splitting/merging patterns over time using a flow metaphor. Later, this work was extended to analyze the complex splitting and merging patterns of hierarchical topics [9] and text streams [25]. Storyline visualization [24, 41, 42] was developed to depict dynamic social interactions in a story or an event, over time. Recently, ThemeDelta [15] was developed to identify significant topic shifts by analyzing how trend keywords converge into topics and diverge into different topics. SocialHelix [7] used DNA helices to visually convey sentiment trends (e.g., divergences and co-occurrences) of online communities. However, these approaches cannot be directly applied to analyze idea flows across multiple groups, as they do not explicitly model the lead-lag relationships on topics. Furthermore, the visualizations employed by these approaches are not favorably situated for comparing multiple idea flows over a large number of time points. Compared

with these approaches, our work focuses on modeling ideas and identifying their lead-lag relationships within/across multiple social groups. We also improve the scalability and flexibility of the exploration by integrating the strengths of a Voronoi-treemap-based bubble tree, a correlated-clustering-based flow map, and a focus+context timeline.

The approach closest to IdeaFlow is TextPioneer [21], which explores the lead-lag relationships among multiple text corpora on hierarchical topics. The differences between our work and TextPioneer are as follows. First, TextPioneer examines lead-lag with regards to one specific topic (e.g., an economy-related topic) across groups, while our system identifies lead-lag relationships between a set of different but correlated ideas (e.g., a healthcare-related idea and a medical-insurance-related idea) within and across groups. Second, TextPioneer only employs word-based content to model the lead-lag relationships and the model accuracy is not satisfactory. Compared with this method, we have developed a correlation model that combines correlations between word time series, the tweet content, and meta-data to generate a more accurate temporal correlation of words. The evaluation in Sec. 6 demonstrates that the model accuracy is significantly improved. Furthermore, we have developed a coordinated visualization that combines a correlated-clustering-based flow map, a focus+context timeline, and a Voronoi-treemap-based bubble tree to illustrate the lead-lag relationships at different levels of granularity.

3 Design of IdeaFlow

Following the nested model described in [30], IdeaFlow was designed through an iterative process with three phases. In the first phase, two domain experts, a professor in media and communications (P1) and a professor in public health (P2), were interviewed regarding their research related to social media. The interviews lasted about one hour each. During the interviews, we learned about the analysis questions, analytic process, and major challenges of concern to the experts. Based on this information, we extracted a set of requirements for IdeaFlow. In the second phase, a prototype based on the requirements was designed, developed, and refined through a set of interviews with the experts. This iterative process took about six months. Alternative designs (such as edge bundling for visualizing the lead-lag relationships) and individual components of the prototype were evaluated by the experts. In the third phase, preliminary case studies were conducted by the visualization experts and presented to the domain experts as a demo. After watching the demo, the domain experts explored their own datasets with the visualization experts using the prototype (P1 spent two hours and P2 spent one hour). Afterwards, the domain experts evaluated the advantages and disadvantages of the system as a whole.

3.1 Analytical Process and Challenges

Analysis questions. P1 and P2 studied political communications and health threats, respectively. In the first phase interview, P1 talked about how he studied the discussions and debates among U.S. congress members, while P2 focused her study on the Ebola outbreak. Both experts claimed that their studies were driven by investigating social issues and their interactions based on who, what, and when, which was referred to by P1 as exploring the Lasswellian questions in communication research [44]. Detailed analysis questions are listed along with the extracted requirements in Section 3.2.

Analytical process and challenges. In their studies, both experts need to extract relevant social issues from the unstructured Twitter data and examine how multiple social issues would interact with one another on Twitter and how such interactions would evolve over time. Moreover, they are interested in identifying influential users and institutions that would initiate the discussion and drive the changes in the discussion of an issue. To implement these tasks in their current practice, the experts use keyword searches to extract relevant tweets, manually code a large number of tweets, and summarize the social issues and track the interactions among relevant issues. This process involves a large amount of human effort and important information may be missed in this labor-intensive manual process. The experts agreed that there was a direct mapping between a "social issue" and an "idea" in our approach and that the lead-lag relationships within and across social groups reflect the interactions and influence of social issues.

3.2 Design Requirements

By abstracting the insights derived from the interview, we distilled the following requirements for the visual analytics system.

R1. Generating an overview of ideas. The experts said they often started their analysis by turning the unstructured raw data into a structured, meaningful overview. The overview consists of descriptions of a set of ideas and their global lead-lag relationships. In particular, the experts identified the following needs:

R1.1. Summarizing ideas and discovering their global lead-lag relationships. In particular, the system should automatically extract meaningful ideas and global lead-lag relationships and present the information to the experts in a manageable manner. This requirement was derived from the analysis questions of the experts such as: "*How many categories of ideas are there in the data? What is their major focus? Which social groups take the lead in these ideas? What are the overall patterns of lead-lag relationships?*" The answers to these types of questions enable experts to gain a comprehensive understanding of the dataset and quickly identify ideas and lead-lag relationships of interest.

R1.2. Examining ideas at different levels of granularity. The experts said in their current practice, it is a natural process to aggregate and disaggregate the ideas. P1 said it is of limited theoretical significance if the system focused on the aggregated ideas (e.g., economy) only. Instead, they were more interested in some concrete sub-ideas (e.g., healthcare, medical insurance) underlying an abstract idea. However, when they analyzed the data, they preferred to start with aggregated ideas first and then decompose high-level patterns to low-level patterns.

R2. Exploring idea flows. Developing a comprehensive understanding of interactions between ideas is among the major concerns of the experts' studies. Accordingly, the following requirements are proposed for exploring correlated ideas and their interactions:

R2.1. Tracking local lead-lag relationships within/across groups on a set of correlated ideas. Correlated ideas would interact with each other, which can enhance social influence of specific social groups on some ideas [31]. During their research, the experts often study analysis questions such as: "*Which social groups took the lead during different time periods such as the weekend / weekdays / voting period / flaring up period? Did the lead-lag relationships change during different phases of an election (or Ebola outbreak) or did they remain stable?* "

R2.2. Exploring and comparing idea flows at different levels of detail (time, social group). First, tracking and comparing local lead-lag relationships over time addresses the aspect of "when" in the Lasswellian questions, allowing the experts to identify global trends of idea flows and interesting lead or lag relationships (e.g., a lead relationship with high importance). After multiple interesting lead or lag relationships are detected, experts often need to decompose and compare them at finer granularity (e.g., by week or by day).

Second, in many applications, social groups are naturally organized in a hierarchy. Accordingly, the experts need to compare idea flows at different levels of the group hierarchy. For example, professor P1 said, "In my current research, I need to compare ideas between Democrats and Republicans. Within each party, I am also curious about the flow behaviors between House Representatives and Senators."

R2.3. Identifying influential social groups. Identifying influential

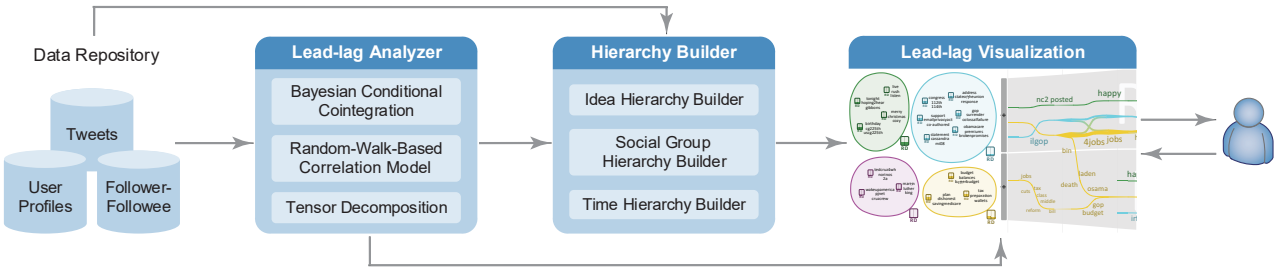

Figure 2: System overview.

user groups in the flow of different ideas is very significant in both experts' research. It allows them to study the social impact and communication strategy of the user groups. Typical analysis questions studied by the experts are: "*Which social groups took the lead during this time period regarding a specific idea? Overall, which social groups are most influential?*"

R3. Exploring details. When describing their analytic process, both experts emphasized the importance of accessing and exploring source data in full detail. Such a detailed examination is important for pattern exploration and hypothesis development and testing. Accordingly, the system should support:

R3.1. Exploring the key content. To understand an idea or a lead-lag relationship, it is critical to access the key content, including the tweets and keywords to depict the idea or the lead/lag behavior. This requirement was derived from analysis questions such as: "*Are there any tweets that can indicate the lead-lag relationships? Which keywords are associated with the lead-lag change?*"

R3.2. Identifying the key users and understanding their roles. The experts also said identifying the key users who drove the evolution of lead-lag relationships is one of their most important tasks, as it can allow policymakers and industrial practitioners (e.g., PR professionals) to take appropriate action at niche time points to control or expedite the evolution. Typical analysis questions are: "*Who are the key players in this idea? Who played the key role in this lead-lag relationship?*"

3.3 System Overview

Based on the aforementioned requirements, we have developed IdeaFlow, which consists of three modules: a lead-lag analyzer, a hierarchy builder, and lead-lag visualization (Fig. 2). Given a set of tweets and the corresponding meta-data, the lead-lag analyzer models ideas and their lead-lag relationships within/across social groups. To enable users to examine the ideas and their lead-lag relationships at different levels of granularity, the hierarchy builder constructs hierarchies for ideas, social groups and time. The analysis results and the hierarchies are then passed to the lead-lag visualization module, which contains three major visual components. A bubble tree is designed to display the overall idea structure (R1.1, R1.2). A correlated-clustering-based flow map visualization is developed to depict the correlated ideas and their lead-lag relationships within and across groups over time (R2.1, R2.3). It is coupled with a multi-focus timeline to enable a flexible exploration of the time dimension (R2.2). The system also provides several interactions for exploring and comparing detailed information (R3.1, R3.2).

4 Tracking Idea Flows

In this section, we introduce our algorithm to detect ideas and track their flows within and across multiple social groups.

4.1 Basic Idea

To model idea flows within and across multiple groups, we need to derive lead-lag relationships between ideas from multiple groups. In Fig. 3(b), solid lines represent lead-lag relationships across different groups, while dashed lines represent lead-lag relationships within the same group. As a result, the key is to learn the lead-lag relationships between ideas. Since each idea is described by a set of words, lead-lag relationships between ideas can be derived by aggregating the temporal correlations between words. Based on this observation, we develop an idea flow tracking algorithm that consists of the following steps.

- **Augmented word graph construction**. In this step, we build an augmented word graph by extracting temporal correlations between words. As shown in Fig. 3(a), each vertex in the graph represents a word from a specific group. Compared with a traditional graph, the edge in the augmented graph encodes the temporal correlations between two words, which is represented by a correlation vector $\boldsymbol{c} = [c_1, ..., c_T]$ and a lead-lag vector $\boldsymbol{\Delta t} = [\Delta t_1, ..., \Delta t_T]$.
- **Identification of ideas and their flows**. In this step, ideas are derived by partitioning the augmented word graph (Fig. 3(b)). Then we aggregate the temporal correlations between words into lead-lag relationships between ideas.

4.2 Augmented Word Graph Construction

Given N social groups and M words, each word from a specific social group is regarded as a vertex v_n^m $(1 \le n \le N, 1 \le m \le M)$ in the augmented graph. Accordingly, the goal of augmented word graph construction is to derive accurate temporal correlations between two vertices. As shown in Fig. 3, the temporal correlations consist of a correlation vector $\boldsymbol{c} = [c_1, ..., c_T]$ and a lead-lag vector $\boldsymbol{\Delta t} = [\Delta t_1, ..., \Delta t_T]$. Here, T is the number of time points, c_k denotes the correlations between the vertices at the kth time point, and Δt_k represents the lead-lag time between the vertices at the kth time point.

The key challenge is to accurately calculate the temporal correlations between words. To this end, we have developed a **random-walk-based correlation model** that takes into account tweet count, meta-data, and stationary behavior between word time series.

Our model is inspired by a three-level mutual reinforcement model [20] used in information retrieval. As shown in Fig. 4, the correlation model consists of three graphs (i.e., user graph, tweet graph, and word graph) and the links between the graphs. The user graph, tweet graph, and word graph are built based on the follower-followee relationships, retweets, and cointegration [5] between word time series, respectively. The links between the graphs

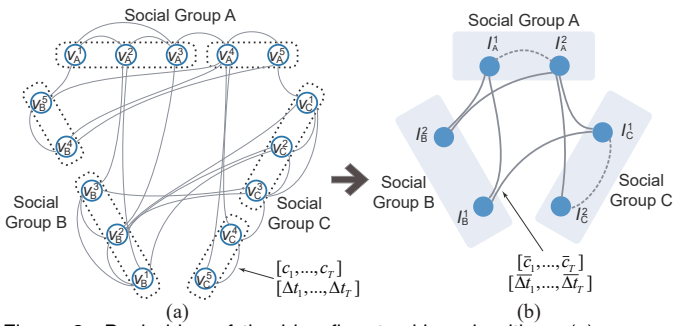

Figure 3: Basic idea of the idea flow tracking algorithm: (a) augmented word graph; (b) lead-lag relationships between ideas.

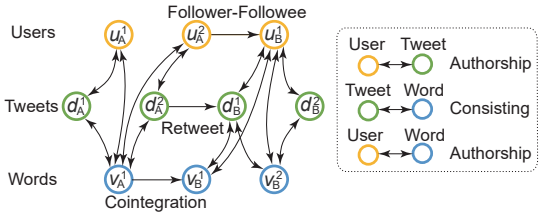

Figure 4: Three-level random-walk-based model for temporal correlation calculation.

are generated based on the tweet content and authorship. Specifically, we link each user to the tweets s/he posted and the words s/he mentioned (authorship). Each tweet is also linked to the words it contains (tweet content). At each time k, we build a correlation model based on the tweets posted during $[k, k+\tau]$. Here τ is the maximum allowable lead-lag time.

Next, we introduce how to derive c_k and Δt_k based on the correlation model. Suppose we need to calculate c_k and Δt_k between v_A^1 and v_B^1. If v_A^1 and v_B^1 are correlated at the kth time point, a short random walk that starts from v_A^1 should reach v_B^1 frequently [34]. Based on this idea, our method is divided into three steps.

First, we perform a series of random walks on v_A^1. A random walk is a path that consists of a succession of random steps. In each step, a random walk may stop at a probability of p_s or move to a neighboring item (i.e., a user, a word, or a tweet) with a probability proportional to the edge weight. For example, in Fig. 4, a possible random walk is $\{v_A^1 \to d_A^2 \to d_B^1 \to v_B^1\}$.

Second, we filter invalid random walks. Not all random walks are valid in our application. For example, if d_A^2 is posted after d_B^1, the information of d_A^2 cannot flow to d_B^1. In this case, $\{v_A^1 \to d_A^2 \to d_B^1 \to v_B^1\}$ is invalid. Formally, we say a random walk l is valid if and only if $t(d^{(l_1)}) \le t(d^{(l_2)}) \le \ldots \le t(d^{(l_Q)})$. Here Q is the number of tweets in l, $d^{(l_q)} (1 \le q \le Q)$ is the qth tweet in l, and $t(d^{(l_q)})$ denotes the time when $d^{(l_q)}$ was posted.

Third, we calculate c_k and Δt_k based on the valid random walks. Specifically, c_k is the empirical probability that a valid random walk starting from v_A^1 visits v_B^1:

$$c_k = |\mathcal{L}(v_A^1 \to v_B^1)| / |\mathcal{L}(v_A^1)|, \tag{1}$$

where $\mathcal{L}(v_A^1)$ is the set of valid random walks that start from v_A^1, $\mathcal{L}(v_A^1 \to v_B^1)$ is the set of valid random walks that start from v_A^1 and reach v_B^1, and $|\cdot|$ denotes the cardinality of a set.

We then calculate Δt_k by averaging the lead-lag time inferred from valid random walks:

$$\Delta t_k = \frac{\sum_{l \in \mathcal{L}(v_A^1 \to v_B^1)} \left[t(d^{(l_{Q'})}) - t(d^{(l_1)}) \right]}{|\mathcal{L}(v_A^1 \to v_B^1)|}, \tag{2}$$

where $d^{(l_{Q'})}$ is the last tweet before v_B^1 in random walk l and $d^{(l_1)}$ is the first tweet in random walk l.

4.3 Identification of Ideas and Their Flows

The major challenge in identifying ideas and their flows is to effectively partition the augmented word graph. Traditional graph partition methods cannot be directly applied to an augmented graph, since each edge in the augmented graph is represented by two vectors instead of a real number. To tackle this issue, we use a tensor to model lead-lag within and across multiple groups jointly.

As shown in Fig. 3(a), the augmented word graph consists of words from multiple social groups and temporal correlations, as represented by two vectors. In order to encode this multi-dimensional data into a consistent model, we first extend the 4-order tensor representation in [53] to a 6-order tensor representation $\boldsymbol{X} \in \mathbb{R}^{N \times M \times N \times M \times T \times (\tau+1)}$. In this representation, the first four dimensions encode the two vertices of an edge, the 5th dimension encodes time, and the 6th dimension encodes the lead-lag time. Specifically, we set $\boldsymbol{X}_{nmn'm'kh}$ to c_k if, at the kth time point, the lead-lag time between vertices v_n^m and $v_{n'}^{m'}$ satisfies $\Delta t_k = h$. Otherwise, we set $\boldsymbol{X}_{nmn'm'kh}$ to 0.

Then we employ a tensor decomposition technique called greedy PARAFAC [18] to derive a feature vector for each word. Finally, we leverage non-negative matrix factorization [51] to detect word clusters based on the feature vectors. Each word cluster is an idea. The correlation between two ideas is calculated by summing up correlations between words: $\bar{\boldsymbol{c}} = \sum \boldsymbol{c}$. The lead-lag time between two ideas at the kth time point is calculated by the weighted average of lead-lag time between words at the kth time point: $\overline{\Delta t_k} = \sum(c_k \Delta t_k) / \sum c_k$.

5 Visualization

According to the requirements discussed in Sec. 3.2, we have developed a visualization that combines the strengths of a Voronoi-treemap-based bubble tree, a flow map, and a timeline.

5.1 Idea View

To summarize a large number of ideas (R1.1) and facilitate their exploration at multiple levels of granularity (R1.2), we organize the ideas into a hierarchy by using the Bayesian Rose Tree model [37]. For each idea (or idea cluster) in the hierarchy, we extract its global lead-lag relationships (R1.2). Next, we introduce the visualization techniques developed to illustrate the global lead-lag relationships and idea hierarchy.

Global lead-lag as glyph. Based on the requirements gleaned from the interview, we represent each idea as meaningful keywords to summarize the core content. Accordingly, the global lead-lag relationships should be placed close to the keywords to facilitate efficient correlation [47]. Inspired by the word-sized sparkline visualization [4], we utilize a word-sized glyph to coherently represent idea keywords and associated global lead-lag relationships. Our design is shown in Fig. 5. In this design, each social group is represented by a word-sized bar chart. The height of the colored bar encodes the proportion of time led by the corresponding social group on the specific idea. Specifically, the height of the colored bar is $h_b T_l / T$, where h_b is the total bar height (Fig. 5), T_l is the number of time points led by the social group on the specific idea, and T is the total number of time points in the dataset.

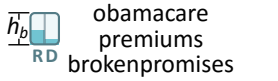

Figure 5: Global lead-lag glyph.

Idea hierarchy as Voronoi-treemap-based bubble tree. The experts were interested in both idea clusters (R1.2) and significant ideas in the clusters (R1.1). Thus we extract representative ideas for each cluster and display their keywords and global lead-lag relationships along with the high-level clusters. The top keywords of each idea cluster are displayed at the bottom of the idea view (Fig. 1(a)). The representative ideas are selected by using topic ranking techniques developed in TIARA [26]. In order to summarize the idea clusters with a substantial number of representative ideas, a space-efficient representation of the idea hierarchy is desirable. A straightforward choice is a space-filling visualization such as Voronoi treemap [3]. However, directly applying the Voronoi treemap would result in difficulties in the global lead-lag glyph layout, since the shapes of the nodes in the Voronoi treemap are often irregular. Moreover, the boundaries of clusters in the Voronoi treemap are not very smooth, which hinders the readability of the clusters [36]. To solve these problems, we combined the Voronoi treemap with EulerSmooth [36] to generate a space-efficient bubble tree as shown in Fig. 1(a).

Layout. To layout the bubble tree, we first use Voronoi treemap to generate a compact layout of the representative ideas and the idea clusters at the user-selected granularity. Then we refine the layout by using EulerSmooth. EulerSmooth is a force-directed method for boundary improvement of Euler diagrams. We chose EulerSmooth because it is able to 1) ensure ideas (represented by keywords and

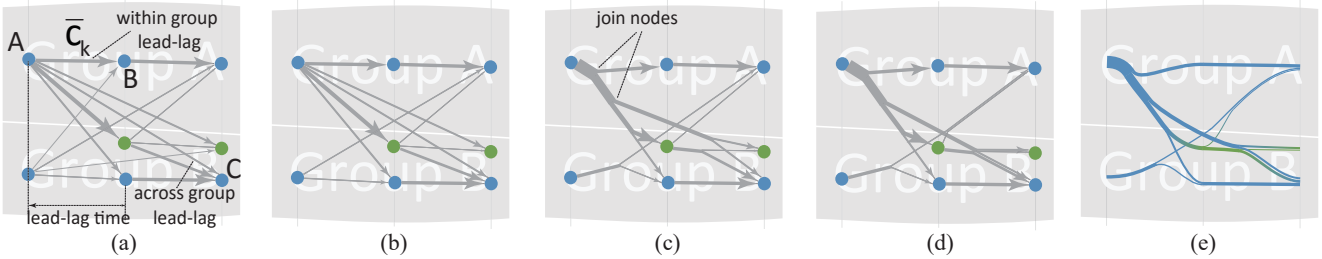

Figure 6: Correlated-clustering-based flow map layout: (a) local lead-lag relationships formulated as a DAG; (b) DAG after applying ambiguity-aware filtering; (c) flow maps generated by clustering target nodes of different source nodes independently; (d) flow maps generated by using correlated clustering; (e) final flow map visualization after smoothing.

global lead-lag glyphs) do not overlap with each other; 2) generate smoothed, aesthetically pleasing boundaries for idea clusters; and 3) ensure ideas remain inside the boundaries of their idea clusters during the refinement. After the refinement, we place the global lead-lag glyphs of idea clusters beside the smoothed boundaries, and refine the layout again by using EulerSmooth.

Interaction. The bubble tree empowers users to examine the ideas at multiple levels of granularity (R1.2). A user can select an idea cluster to zoom in or click the zoom out button (Fig. 1E) to return to high-level views. Staged animation [2] is used to preserve a user's mental map during the zoom process. We also employ a polygon-clipping algorithm [45] to better show how idea clusters split into sub-clusters or how sub-clusters merge during the animation.

5.2 Flow View

Fig. 1(b) shows a flow map view that combines several visualizations to address the exploration requirements of the domain experts (Sec. 3.2). Specifically, we encode multi-level user groups as a stripe tree (R2.2, R2.3) and local lead-lag as flow maps [6] (R2.1). The flow map is coordinated with a focus+context timeline, which enables multi-resolution exploration of time (R2.2).

Multi-level social group as stripe tree. Fig. 1(b) shows a stripe tree, which consists of a stacked tree [49] and user group stripes. Each node in the stacked tree represents a social group. For each leaf node in the stacked tree, we display its activeness over time by using a stripe. The activeness is defined as the total number of tweets posted by the users in this group at the specific time point.

In our first implementation, we place social group names (e.g., "Republicans") on the left side of the stripes. When we presented this implementation to P1, he felt it was difficult to recognize social group names when he observed idea flows. He said he had to move his eyes back and forth to check for social group names when he tried to understand the idea flows. Even worse, he might confuse the group names with the content keywords on each stripe. Following his suggestion, we implemented a watermark design (Fig. 1G) and showed both implementations to 12 potential users. Nine of them preferred the watermark design and all of them agreed that the watermark does not influence their exploration of the idea flows because of enough contrast between the white watermark and the keywords. Thus we adopted the watermark design in our final system.

Interaction. The stacked tree enables users to explore social groups at multiple levels of granularity (R2.2). Users can click a leaf node of the stacked tree to zoom into lower level social groups or click an intermediate node to zoom out. We use staged animation to smooth the zoom in/out transition and preserve users' mental maps.

Local lead-lag as flow map. Mathematically, local lead-lag relationships between ideas can be formulated as a Directed Acyclic Graph (DAG). As shown in Fig. 6(a), each node in the DAG represents an idea from a social group at a specific time. The nodes are vertically ordered and placed by using an algorithm developed by Xu et al. [50]. The color of the node represents the idea cluster it belongs to. Two ideas are connected by a directed edge if one leads the other. In this way, the DAG encodes within group lead-lag (e.g., edge $\overrightarrow{AB}$), across group lead-lag (e.g., $\overrightarrow{AC}$), and lead-lag time. Each edge is assigned a weight (e.g., $\bar{c}_k$ on edge $\overrightarrow{AB}$) that encodes the correlation between corresponding ideas from multiple groups.

A straightforward way to visualize the DAG is to use Sankey diagrams [33], a widely-used method for visualizing weighted information flow. However, Sankey diagrams may cause visual clutter when they are used to display flows between many categories [49]. Another option is to use edge bundling [16] to merge the flows and reduce clutter. However, edge bundling may cause many crossings and thus introduce unnecessary visual clutter [6]. To solve the problem, we use flow maps since they can effectively reduce visual clutter by merging edges quickly and smoothly [6, 10].

Layout. Flow maps are designed to analyze the movement of objects from one source location to multiple target locations [32]. Each flow map contains one source node and multiple target nodes. Since there are many source nodes (i.e., nodes with outgoing edges) in the DAG, we need to layout many flow maps, which may result in visual clutter and ambiguity. To effectively layout multiple flow maps simultaneously and avoid ambiguity, we have developed a correlated-clustering-based flow map layout algorithm that consists of the following steps.

The first step is **ambiguity-aware filtering**. In this step, we reduce ambiguity of the DAG by filtering the less important edges. Importance is measured by the edge weight $\bar{c}_k$. To measure ambiguity, we employ four metrics proposed in AmbiguityVis [46], namely node occlusion (g_k^o), edge crossing (g_k^c), edge crossing angle (g_k^a), and node-edge occlusion (g_k^n). For each edge, we use the four metrics to calculate g_k, which is the amount of ambiguity caused by this edge. Specifically, g_k is defined as $\mu_o g_k^o + \mu_c g_k^c + \mu_a g_k^a + \mu_n g_k^n$, where μ_o, μ_c, μ_a, and μ_n are parameters that balance the four ambiguity metrics. In IdeaFlow, we set $\mu_o = \mu_c = \mu_n = 1$ and $\mu_a = 0.01$. We filter the kth edge if $\bar{c}_k \leq \gamma g_k$. Here $\bar{c}_k$ is the weight of the edge and γ is a user-provided parameter that balances ambiguity and the amount of information provided. The default value of γ is set to 0.02, which we find generates acceptable layout results in many cases. We also allow users to adjust γ based on their information needs by using a slider (Fig. 1J). Fig. 6(b) shows the DAG after applying ambiguity-aware filtering.

The second step is **correlated clustering.** In this step, we calculate the join nodes of the flow maps. Join nodes are nodes where the edges merge (Fig. 6(c)). A state-of-the-art algorithm [6] calculates join nodes by clustering the target nodes (e.g., Fig. 6C) into a spiral tree. When there are multiple source nodes, this algorithm clusters target nodes of each source node independently. As shown in Fig. 6(c), while this algorithm does reduce the visual clutter of the DAG, it may fail to merge adjacent edges of different source nodes, causing unnecessary visual clutter. To solve this problem, we use correlated clustering [17] to simultaneously build multiple spiral trees. As shown in Fig. 6(d), our algorithm is able to detect similar structures that belong to different flow maps and merge corresponding edges to further reduce visual clutter.

The final step is **smoothing**. In this step, we connect the join

nodes and target nodes by using cubic Bezier curves. The control points of the cubic Bezier curves are placed so that the derivative of the incoming curve and outgoing curve are the same at the place where they intersect. This results in a smoother transition across different flow maps (Fig. 6(e)).

Focus+context timeline as spring. The stripe tree and flow map are coupled with a multi-focus timeline, which aims to facilitate the comparison and exploration of idea flows at multiple levels of time granularity (R2.2). To effectively convey the compression of time, we use a spring metaphor inspired by thread weaving in TextFlow [8]. Specifically, the spring metaphor encodes the number of time points compressed in the frequency of the spring (Fig. 1(b)).

Interaction. Users can select a time period by clicking the timeline, then expand or shrink it by dragging the left or right side of the time period. During the dragging process, the flow map visualization and user group stripes are updated in real-time. If enough space is given, the time period will split into multiple time periods with a smaller time granularity. User action on the timeline will be automatically recorded and displayed in the timeline action trail panel (Fig. 1H), which enables users to label the time period and easily return to previous time periods (R2.2).

5.3 Information Panel

To facilitate the exploration of the source data in full detail (R3), we designed an information panel that consists of four components: a word panel, a user panel, a tweet panel, and a timeline action trail panel. Users can click an idea flow to observe its key content in the word and tweet panel (R3.1) and identify the key users by observing the user panel (R3.2).

6 Evaluation

In this section, a quantitative evaluation and two case studies were conducted to demonstrate the usefulness and effectiveness of IdeaFlow. The following two Twitter datasets were used in the evaluation:

- **Congress dataset A**, which contains 1,605,361 tweets posted by 1,102 Twitter accounts from the members of the 114th U.S. Congress (January 2010 to March 2016). These accounts were manually identified and labeled by a master's student of communications. We classified the accounts into four groups: Democratic or Republican Senators and Democratic or Republican House Representatives.
- **Ebola dataset B**, which contains 16,711,670 tweets posted by 321,114 accounts from July 2014 to February 2016. We grouped the accounts according to the location information in Twitter profiles. Locations were cleaned and organized into a hierarchy using a location enrichment API [55].

The datasets were processed using the following three steps. First, we grouped the tweets into different segments, each of which contained one day of tweets. In total, we got 2,253 segments (time points) for Dataset A and 574 segments (time points) for Dataset B. Second, we built a correlation model (Fig. 4) for each time point k based on kth to $(k+\tau)$th segments and used this model to calculate c_k and Δt_k. Here we set τ to 3. In our experience, this setting well balances model accuracy and efficiency. Third, we constructed the augmented word graph by using c_k and Δt_k and used tensor-based techniques (Sec. 4.3) to identify the ideas and lead-lag relationships. Overall, it took 17 hours to process Dataset A and 9 hours to process Dataset B on a workstation with an Intel Xeon E52630 CPU (2.4 GHz) and 64GB Memory. We extracted 300 ideas and 27,802 lead-lag relationships from Dataset A and 100 ideas and 2,568 lead-lag relationships from Dataset B.

6.1 Quantitative Evaluation

In this experiment, we demonstrated that by taking into account tweet content, meta-data, and correlation between time series, the model accuracy was substantially improved.

Baselines. We compared our algorithm with two baselines. The first baseline (**B1**) was the method developed by Zhong et al. [53]. This method uses cointegration to track idea flows and is limited to two groups. In order to compare it with our method, we substituted their tensor representation with ours, so that it could track lead-lag changes within and across multiple groups. The second baseline (**B2**) was similar to our method. The only difference was that it does not take into account the cointegration between word time series. We designed B2 to examine whether model accuracy will be improved by adding cointegration information.

Experimental settings. In the experiment, we measured two kinds of model accuracy: idea accuracy and lead-lag relationship accuracy. Here, accuracy is defined as the proportion of ideas/flows that are labeled to be correct [56]. We invited two PhD students majoring in data mining to label whether the ideas and lead-lag relationships were correct. Both students have experience in data labeling and were familiar with the datasets. For each idea, we provided the students with the top 10 keywords and tweets regarding this idea. An idea is labeled to be correct if most of the keywords can be grouped together as a single coherent thought or opinion. The top tweets were provided to help them understand the idea. For each lead-lag relationship, which contains two ideas, we provided the students with the top keywords of the two ideas and tweet pairs that could potentially explain the lead-lag relationships between the ideas. Since the number of lead-lag relationships was very large ($> 1{,}000$), we asked them to label the 150 most important ones. The inter-annotator agreement between the PhD students was 80.8%. To further improve the labeling quality, we hired a professional coder to investigate the labeling results and correct potential errors.

Results. As shown in Table 1, our method outperformed state-of-the-art method from Zhong et al. [53] (B1), with an idea accuracy improvement of 0.2 and a lead-lag relationship accuracy improvement of 0.165. These results demonstrate that considering tweet content and meta-data improves model accuracy. Moreover, our method also outperformed B2, indicating that taking into account cointegration between word time series also improves model accuracy. Overall, these results demonstrate that the model accuracy is substantially improved by taking into account tweet content, meta-data, and cointegration between time series.

6.2 Case Study

6.2.1 Discussions and Debates among Congress Members

This first case study was conducted with a professor of media and communications (P1). The professor wanted to understand the communicative strategies employed by political parties as a whole and by individual political figures in their interactions on social media. He also expected that the visual analytic system would help him develop more insights into the roles of various social groups on the interplay of social issues. We used Dataset A in this case study.

Obtaining an overview of ideas and their flows (R1.1 and R2.1). To understand what ideas the congress members discussed on Twitter, the professor started by observing the top-level idea clusters (Fig. 1(a)). He immediately identified four idea clusters. They were the economy (yellow), the president and congress (blue), social and cultural events such as holidays (green), and Ted Cruz, a Republican nominee for the President of the United States in the forthcoming 2016 election (purple). Curious about how the ideas flowed among congress members, the professor examined the flow view (Fig. 1(b)). He noticed that the flows of different idea clusters demonstrated very different patterns. For example, Ted-Cruz-

Table 1: Comparison of the model accuracy of different methods.

	Idea			Lead-lag		
	B1	B2	Ours	B1	B2	Ours
Dataset A	0.543	0.753	**0.793**	0.653	0.700	**0.800**
Dataset B	0.790	0.830	**0.940**	0.620	0.800	**0.803**

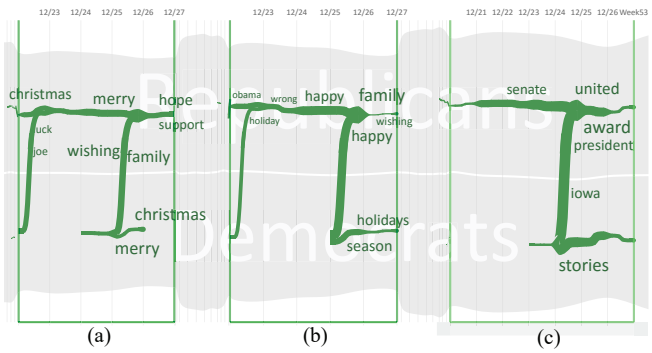

Figure 7: Holiday related ideas.

related ideas (purple) only circulated within one specific social group: Republicans. The flow of social and cultural events (green), whose strength grew over time, was independent from the flow of other idea clusters. The professor also observed that economy-related ideas (yellow) and president- and congress-related ideas (blue) frequently interacted with each other (e.g., **A**), indicating that there was mutual influence between different idea clusters. The professor commented that such interactions make sense because bills that pertain to spending and other economic issues are debated and voted on Congress.

Exploring details of ideas of interest (R3.1 and R3.2). The Ted Cruz-related idea cluster caught the professor's attention, since it only circulated among Republicans (**B**). By exploring more details in the information panel (Fig. 1(c)), the professor found that this idea cluster was generated because Ted Cruz created his own hashtags and posted many tweets about himself in order to enhance his visibility in horse races (**C**). However, his strategy was not very successful: this idea cluster only flowed among his own accounts (**D**).

Zooming into idea clusters of interest and comparing flows from different time periods (R1.2 and R2.2). The professor was curious about non-political ideas in the discussion of the Congress members, and thus he zoomed into the green idea by selecting the largest sub-idea at each level of detail. After zooming into the bottom level, he found five ideas all about different holidays, such as Christmas, Mother's Day, and Easter. To understand why there were so many holiday-related ideas, the professor studied the flow of Christmas-related ideas. Figs. 7(a), 7(b), and 7(c) show the flow in 2013, 2014, and 2015, respectively. The professor noticed that the flow was always stronger among Republicans than among Democrats during the Christmas season. After examining the details, the professor found that the Republicans tended to leverage Christmas for political purposes. For example, during Christmas 2014 (Fig. 7(b)), Republicans mentioned words like "troops," "kenya," and "obama." Related tweets show that because President Obama sent American troops to Kenya, the soldiers were not able to return home for Christmas. The Republicans took advantage of this opportunity to criticize the decision by Barack Obama with tweets such as "Obama didn't go to Kenya for Christmas. RT @Rachel-Bynum: @HavanaTed @TeaPainUSA Going home for Christmas, what is wrong with that?" This created a public relations crisis for the Democrats who were supportive of Barack Obama.

Intrigued by the findings, the professor examined other holidays such as Independence Day. He discovered that the Republicans again tended to leverage this holiday for political purposes, which he referred to as the politicization of holidays. The expert commented, "The politicization has been widely recognized by social researchers and the general public [29]. However, it is not well studied in empirical research. These findings provide solid empirical evidence for the existence of politicization of festivals and culture in the U.S. Congress."

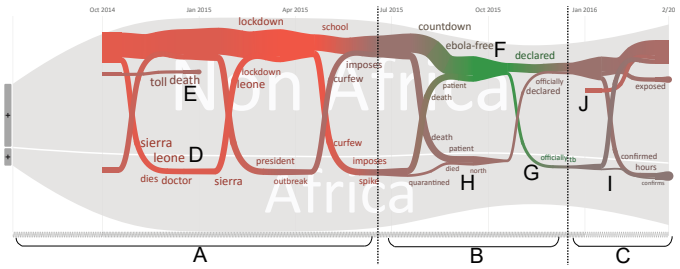

Figure 8: Flow of sentiment towards Sierra Leone related ideas.

6.2.2 The Ebola Epidemic

The widest spread Ebola epidemic started in December 2013. This epidemic had 28,639 reported cases with 11,316 deaths as of Mar. 16th, 2016 [54]. In this case study, we collaborated with professor P2. She is interested in understanding how ideas with different sentiment, flow on social media during a health crisis like Ebola. With this comprehensive understanding, she hoped to make suggestions to governments and intergovernmental organizations to provide suitable responses during an urgent situation. The dataset used here was Dataset B. To meet the professor's analytical needs, we calculated the sentiment of the flow by using a word-embedding-based sentiment classification method [43]. To represent the sentiment, we used a color map that ranges from red (negative sentiment) to gray (neutral sentiment) and green (positive sentiment).

Comparing sentiment flows at different phases of the epidemic (R2.1 and R2.3). P2 was concerned with the African countries severely affected by Ebola. Thus she searched for "sierra leone," which is the name of the African country with the largest number of Ebola cases, and zoomed into the corresponding idea cluster. The flows of this idea cluster are shown in Fig. 8. By observing the sentiment flows and related labels, the professor identified three phases of the epidemic in Sierra Leone.

The **first phase** was the outbreak phase (**A**). During this phase, strong negative sentiment flowed within and across African countries and non-African countries. Some negative reports and messages such as regarding the death of doctors (**D**, "Sierra Leone's Leading Doctor Dies of #Ebola Before ZMab Drug Arrives @WHO") and death toll (**E**) grabbed public attention. The **second phase** was the descending phase (**B**). During this phase, WHO declared that Sierra Leone was Ebola free and strong positive sentiment started to flow within non-African countries (**F**). This positive sentiment influenced (led) the African countries to some extend (**G**). However, negative sentiment continued to flow within African countries ("Sierra Leone is Ebola free but a legacy of fear remains."), which also influenced (led) non-African countries (**H**). The **third phase** was the flaring up phase (**C**). During this phase, a new Ebola case was confirmed hours after the WHO declared the outbreak was over (**I**, "Ebola resurfaces in Sierra Leone hours after WHO declares outbreak over http://cnn.it/233OoMT"). As shown in the flow view, the sentiment of non-African countries tended to be more negative than that of African countries (**J**).

Based on the above observations, P2 suggested that governments and intergovernmental organizations should be more cautious during the descending phase. She said that although the Ebola virus was under control during this phase, negative sentiments continued to spread via social media. Thus guiding the public towards a more rational understanding of Ebola is critical in this phase. She then suggested that international organizations such as the WHO avoid absolute statements, emphasize appropriate reservations, and communicate clearly. In particular, she made the following suggestions: 1) emphasize appropriate reservations about reliability if the crisis is still evolving or information is incomplete; 2) share more information, not less; otherwise people may think something significant is being hidden; and 3) use clear, non-technical language appropri-

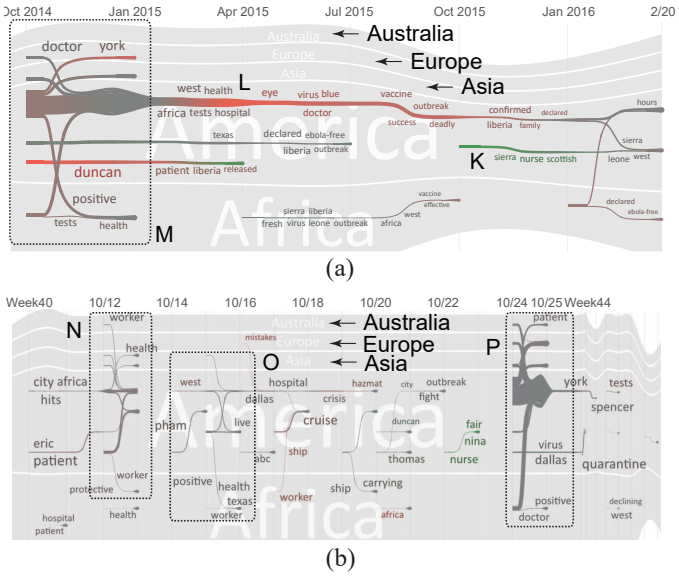

Figure 9: Idea flows within and across multiple continents.

ate to the target audience, for example, stories, narratives, examples, and anecdotes that can be used to make technical data come alive.

Exploring idea flows within and across multiple continents (R2.1 and R2.2). After some exploration, P2 drilled into an idea cluster related to the Ebola cases in the U.S. to study how the Ebola cases outside Africa influenced the public. The flows of this idea cluster are shown in Fig. 9(a). P2 split social group "non-Africa" into sub-groups (e.g. Australia, Europe, Asia, America) to investigate how ideas flow within and across continents. After analyzing idea flows shown in Fig. 9(a), P2 made several interesting observations: 1) Australia was least affected among these continents; 2) strong positive (e.g., **K**) and strong negative (e.g., **L**) sentiment of American users had little influence on the sentiment of users from other continents; and 3) many continents were influenced by these ideas from Oct. 2014 to Jan. 2015 (**M**), when the first Ebola cases in the U.S. were reported ("Texas nurse tests positive for Ebola, would be 1st Ebola transmission in U.S.").

Interested in the first Ebola cases in the United States, P2 drilled into time period **M** to examine more details. After iteratively expanding time periods with the largest number of flows, she drilled into the time period when the first three Ebola cases of American healthcare workers were reported (Oct. 11, 2014 to Oct. 25, 2014). Fig. 9(b) shows how ideas flowed within and across multiple continents during this time period. Flows of the three cases were marked as **N**, **O**, and **P**, respectively. P2 was surprised to find that the most influential case was the third one instead of the first one. To explain this pattern, she examined the patients (who), locations (where), and dates (when) of the three cases by reading the tweets and compared the three cases in terms of these aspects. After the comparison, she recognized the major difference between the third case and the first two cases were on the patients. While the patients of the first two cases are nurses from a Texas hospital, the patient of the third case is a doctor from Doctors without Borders, which is a well-known international non-governmental organization. P2 commented that Doctors without Borders was very active in public health issues and it was reasonable that doctors from this organization received more attention compared with other patients.

7 Discussion of Limitations

Although the evaluation demonstrates the effectiveness of IdeaFlow, our system still has several limitations.

First, our idea flow tracking algorithm is an offline method. The computational cost is dominated by the high cost of tensor decomposition. When calculating the tensor decomposition, the highest cost arises from computing the covariance matrix along each dimension [40]. Thus, the overall time complexity is $O(N^{M+1})$, where M is the dimension number of the tensor and N is the average element number along each dimension. The algorithm usually requires hours of computation time on datasets with millions of tweets. As a result, it does not support interactive parameter adjustment with real-time feedback and is not able to process data in real time as new tweets are streaming in. A possible step towards this goal is to incorporate online tensor decomposition to incrementally update the ideas (and subsequently the lead-lag relationships).

Second, the number of ideas needs to be set manually, which limits the portability of our method to different datasets. This problem can be solved by using measures such as Akaike Information Criterion [1] to automatically determine the number of ideas.

Third, the number of idea clusters that can be clearly presented is limited. This is because we use color to help users find the correspondence between idea clusters on the bubble tree and the flows. We addressed this problem using labels and interactions. In particular, users can find the correspondence between idea clusters and flows by reading the labels in the idea/flow view or hovering over an idea cluster to highlight the corresponding flow and vice versa.

8 Conclusions and Future Work

In this paper, we have presented IdeaFlow, an interactive visual analytics system for exploring lead-lag relationships within and across multiple social groups on a set of correlated ideas. IdeaFlow automatically models local lead-lag changes within and across multiple groups on a set of different but correlated ideas. Furthermore, a visualization consisting of a bubble tree, a novel flow map, and a focus+context timeline is provided to present the model outputs at different levels of time granularity. A quantitative evaluation and two case studies with domain experts have demonstrated the effectiveness and usefulness of IdeaFlow.

Expert feedback has also illuminated future directions for improving our system. According to the domain experts, their data sometimes lacks meta-attributes (e.g., location) that are meaningful for categorizing social groups. Thus, the first area of possible improvements is to develop a method for automatically discovering social groups through semi-supervised or unsupervised learning. Second, we are interested in comparing the ambiguity (e.g., node occlusion, edge crossing, edge crossing angle, and node-edge occlusion) of the developed flow map layout method with previous methods to verify the effectiveness of our visualization. Third, experts expressed the need to perform hypothesis testing on IdeaFlow. For example, they want to change the lead-lag relationship between social groups for an idea at a specific time period, and then they expect to see how the overall picture changes. To meet this need, we plan to investigate possible solutions to support hypothesis testing. One of the domain experts wanted to use IdeaFlow to explore idea flows on multiple corpora, so another interesting venue for future work would be to conduct a case study on multiple textual corpora.

Acknowledgements

This work is supported by National NSF of China, a Microsoft Research Fund (No. FY15-RES-OPP-112), and the National Science Foundation under award number 0915528. The authors would like to thank Mengchen Liu for his valuable suggestions.

References

[1] H. Akaike. *Akaike's Information Criterion*, pages 25–25. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[2] B. Bach, E. Pietriga, and J.-D. Fekete. GraphDiaries: animated transitions andtemporal navigation for dynamic networks. *IEEE TVCG*, 20(5):740–754, 2014.

[3] M. Balzer and O. Deussen. Voronoi treemaps. In *InfoVis*, pages 49–56, 2005.

[4] F. Beck, S. Koch, and D. Weiskopf. Visual analysis and dissemination of scientific literature collections with SurVis. *IEEE TVCG*, 22(1):180–189, 2016.

[5] C. Bracegirdle and D. Barber. Bayesian conditional cointegration. In *ICML*, pages 1095–1102, 2012.

[6] K. Buchin, B. Speckmann, and K. Verbeek. Flow map layout via spiral trees. *IEEE TVCG*, 17(12):2536–2544, 2011.

[7] N. Cao, L. Lu, Y.-R. Lin, F. Wang, and Z. Wen. SocialHelix: visual analysis of sentiment divergence in social media. *Journal of Visualization*, 18(2):221–235, 2015.

[8] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. J. Gao, H. Qu, and X. Tong. TextFlow: towards better understanding of evolving topics in text. *IEEE TVCG*, 17(12):2412–2421, 2011.

[9] W. Cui, S. Liu, Z. Wu, and H. Wei. How hierarchical topics evolve in large text corpora. *IEEE TVCG*, 20(12):2281–2290, 2014.

[10] A. Debiasi, B. Simões, and R. De Amicis. Supervised force directed algorithm for the generation of flow maps. In *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, pages 193–202, 2014.

[11] M. Dörk, D. Gruen, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. *IEEE TVCG*, 16(6):1129–1138, 2010.

[12] W. Dou, X. Wang, R. Chang, and W. Ribarsky. ParallelTopics: a probabilistic approach to exploring document collections. In *IEEE VAST*, pages 231–240, 2011.

[13] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou. LeadLine: interactive visual analysis of text data through event identification and exploration. In *IEEE VAST*, pages 93–102, 2012.

[14] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. HierarchicalTopics: visually exploring large text collections using topic hierarchies. *IEEE TVCG*, 19(12):2002–2011, 2013.

[15] S. Gad, W. Javed, S. Ghani, N. Elmqvist, T. Ewing, K. N. Hampton, and N. Ramakrishnan. ThemeDelta: dynamic segmentations over temporal topic models. *IEEE TVCG*, 21(5):672–685, 2015.

[16] D. Holten and J. J. van Wijk. Force-directed edge bundling for graph visualization. In *EuroVis*, pages 983–998, 2009.

[17] P. Kirk, J. E. Griffin, R. S. Savage, Z. Ghahramani, and D. L. Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297, 2012.

[18] T. Kolda, B. Bader, and J. Kenny. Higher-order web link analysis using multilinear algebra. In *ICDM*, pages 242–249, 2005.

[19] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD*, pages 497–506, 2009.

[20] M. Liu, S. Liu, X. Zhu, Q. Liao, F. Wei, and S. Pan. An uncertainty-aware approach for exploratory microblog retrieval. *IEEE TVCG*, 22(1):250–259, 2016.

[21] S. Liu, Y. Chen, H. Wei, J. Yang, K. Zhou, and S. M. Drucker. Exploring topical lead-lag across corpora. *TKDE*, 27(1):115–129, 2015.

[22] S. Liu, W. Cui, Y. Wu, and M. Liu. A survey on information visualization: recent advances and challenges. *The Visual Computer*, 30(12):1373–1393, 2014.

[23] S. Liu, X. Wang, Y. Song, and B. Guo. Evolutionary bayesian rose trees. *TKDE*, 27(6):1533–1546, 2015.

[24] S. Liu, Y. Wu, E. Wei, M. Liu, and Y. Liu. StoryFlow: tracking the evolution of stories. *IEEE TVCG*, 19(12):2436–2445, 2013.

[25] S. Liu, J. Yin, X. Wang, W. Cui, K. Cao, and J. Pei. Online visual analytics of text streams. *To appear in IEEE TVCG*, 2015.

[26] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian. TIARA: interactive, topic-based visual text summarization and analysis. *ACM TIST*, 3(2):25, 2012.

[27] Y. Lu, M. Steptoe, S. Burke, H. Wang, J. Y. Tsai, H. Davulcu, D. Montgomery, S. R. Corman, and R. Maciejewski. Exploring evolving media discourse through event cueing. *IEEE TVCG*, 22(1):220–229, 2016.

[28] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. Keim. EventRiver: visually exploring text collections with temporal references. *IEEE TVCG*, 18(1):93–105, 2012.

[29] A. M. McCright and R. E. Dunlap. The politicization of climate change and polarization in the American public's views of global warming, 2001–2010. *The Sociological Quarterly*, 52(2):155–194, 2011.

[30] T. Munzner. A nested process model for visualization design and validation. *IEEE TVCG*, 15(6):921–928, 2009.

[31] A. P. Pentland. *Social Physics: how Good Ideas Spread-The Lessons from a New Science*. Penguin Press, 2014.

[32] D. Phan, L. Xiao, R. Yeh, and P. Hanrahan. Flow map layout. In *InfoVis*, pages 219–224, 2005.

[33] P. Riehmann, M. Hanfler, and B. Froehlich. Interactive Sankey diagrams. In *InfoVis*, pages 233–240, 2005.

[34] D. Shahaf and C. Guestrin. Connecting the dots between news articles. In *KDD*, pages 623–632, 2010.

[35] L. Shi, F. Wei, S. Liu, L. Tan, X. Lian, and M. X. Zhou. Understanding text corpora with multiple facets. In *IEEE VAST*, pages 99–106, 2010.

[36] P. Simonetto, D. Archambault, and C. Scheideg. A simple approach for boundary improvement of Euler diagrams. *IEEE TVCG*, 22(1):678–687, 2016.

[37] Y. Song, S. Liu, X. Liu, and H. Wang. Automatic taxonomy construction from keywords via scalable bayesian rose trees. *TKDE*, 27(7):1861–1874, 2015.

[38] G. Sun, Y. Wu, R. Liang, and S. Liu. A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *JCST*, 28(5):852–867, 2013.

[39] G. Sun, Y. Wu, S. Liu, T.-Q. Peng, J. J. Zhu, and R. Liang. EvoRiver: visual analysis of topic coopetition on social media. *IEEE TVCG*, 20(12):1753–1762, 2014.

[40] J. Sun, D. Tao, S. Papadimitriou, P. S. Yu, and C. Faloutsos. Incremental tensor analysis: Theory and applications. *TKDD*, 2(3):11:1–11:37, 2008.

[41] Y. Tanahashi, C.-H. Hsueh, and K.-L. Ma. An efficient framework for generating storyline visualizations from streaming data. *IEEE TVCG*, 21(6):730–742, 2015.

[42] Y. Tanahashi and K.-L. Ma. Design considerations for optimizing storyline visualizations. *IEEE TVCG*, 18(12):2679–2688, 2012.

[43] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning sentiment-specific word embedding for Twitter sentiment classification. In *ACL*, pages 1555–1565, 2014.

[44] L. Van Zoonen. *Feminist media studies*, volume 9. Sage, 1994.

[45] B. R. Vatti. A generic solution to polygon clipping. *Communications of the ACM*, 35(7):56–63, 1992.

[46] Y. Wang, Q. Shen, D. Archambault, Z. Zhou, M. Zhu, S. Yang, and H. Qu. AmbiguityVis: visualization of ambiguity in graph layouts. *IEEE TVCG*, 22(1):359–368, 2016.

[47] C. Ware. *Information visualization: perception for design*. Elsevier, 2012.

[48] F. Wu, Y. Song, S. Liu, Y. Huang, and Z. Liu. Lead-lag analysis via sparse co-projection in correlated text streams. In *CIKM*, pages 2069–2078, 2013.

[49] Y. Wu, S. Liu, K. Yan, M. Liu, and F. Wu. OpinionFlow: visual analysis of opinion diffusion on social media. *IEEE TVCG*, 20(12):1763–1772, 2014.

[50] P. Xu, Y. Wu, E. Wei, T.-Q. Peng, S. Liu, J. J. Zhu, and H. Qu. Visual analysis of topic competition on social media. *IEEE TVCG*, 19(12):2012–2021, 2013.

[51] W. Xu, X. Liu, and Y. Gong. Document clustering based on Non-negative Matrix Factorization. In *ACM SIGIR*, pages 267–273, 2003.

[52] R. Zhao and W. I. Grosky. Bridging the semantic gap in image retrieval. *Distributed multimedia databases: Techniques and applications*, pages 14–36, 2002.

[53] Y. Zhong, S. Liu, X. Wang, J. Xiao, and Y. Song. Tracking idea flows between social groups. In *AAAI*, pages 1436–1443, 2016.

[54] Ebola situation report released by WHO on Mar. 16th, 2016. `http://apps.who.int/ebola/current-situation/ebola-situation-report-16-march-2016`, Mar. 2016.

[55] Location enrichment API provided by FullContact. `https://www.fullcontact.com/developer/docs/location/#location-enrichment`, Mar. 2016.

[56] Definition of accuracy in Wikipedia. `https://en.wikipedia.org/wiki/Accuracy_and_precision`, Mar. 2016.