# ThemeDelta: Dynamic Segmentations over Temporal Topic Models

Samah Gad, *Member, IEEE*, Waqas Javed, *Member, IEEE*, Sohaib Ghani, *Member, IEEE*, Niklas Elmqvist, *Senior Member, IEEE*, Tom Ewing, Keith N. Hampton, and Naren Ramakrishnan, *Member, IEEE*

**Abstract—**We present ThemeDelta, a visual analytics system for extracting and visualizing temporal trends, clustering, and reorganization in time-indexed textual datasets. ThemeDelta is supported by a dynamic temporal segmentation algorithm that integrates with topic modeling algorithms to identify change points where significant shifts in topics occur. This algorithm detects not only the clustering and associations of keywords in a time period, but also their convergence into topics (groups of keywords) that may later diverge into new groups. The visual representation of ThemeDelta uses sinuous, variable-width lines to show this evolution on a timeline, utilizing color for categories, and line width for keyword strength. We demonstrate how interaction with ThemeDelta helps capture the rise and fall of topics by analyzing archives of historical newspapers, of U.S. presidential campaign speeches, and of social messages collected through iNeighbors, a web-based social website. ThemeDelta is evaluated using a qualitative expert user study involving three researchers from rhetoric and history using the historical newspapers corpus.

**Index Terms—**Language models, time-series segmentation, text analytics, visual representations

## 1 Introduction

Several visual analytic applications require the analysis of dynamically changing trends over time. Example contexts include studies of idea diffusion in scientific communities, the ebb and flow of news on global, national, and local levels, and the meandering patterns of communication in social networks. Trends, each representing a particular keyword or concept, that converge into topics at different points in time, then just as unpredictably diverge into new defined topics at a later time, are key patterns of interest to an analyst. Both experts and casual users alike need mechanisms for understanding such evolving trends for analysis, prediction, and decision making.

While much research exists in data mining and visualization, we posit that they are insufficient to address the needs of these emerging applications. Existing visualization techniques such as ThemeRiver [1] and streamgraphs [2] are aimed to capturing overall trends in textual corpora but fail to capture their branching and merging nature. Concomitantly, data mining algorithms have evolved greatly over the past years, especially for topical text modeling [3], but capturing key breakpoints in topic evolution and defining appropriate visual representations for such breakpoints is an understudied problem. In this paper, we present ThemeDelta, a visual analytics system for accurately extracting and portraying how individual trends gather with other trends to form ad hoc groups of trends at specific points in time. Such gathering is inevitably followed by scattering, where trends diverge or fork to form new groupings. Understanding the interplay between these two behaviors provides significant insight into the temporal evolution of a dataset. We built ThemeDelta to support hypothesis generation by focusing on developing a system for exploring and navigating trends. A natural extension to ThemeDelta is expanding the system capabilities to further support robust knowledge generation, but this is beyond the scope of this paper.

There has been significant research in capturing the dynamic evolution of topics underlying text corpora. Most of these efforts are focused on extending the classical probabilistic model of Latent Dirichlet Allocation (LDA) [3] to a time-indexed context. Our ThemeDelta temporal topic modeling approach is differentiated by its emphasis on automatically identifying segments where topic distribution is uniform and segment boundaries around which significant changes are occurring. We embed a temporal segmentation algorithm around a topic modeling algorithm to capture such significant shifts of coverage. A key advantage of our approach is that it integrates with existing topic modeling algorithms in a transparent manner; thus, more sophisticated algorithms can be readily plugged in as research in topic modeling evolves.

Similar to how ThemeRiver [1] uses a single stacked graph centered around a horizontal timeline that resembles a physical river, our ThemeDelta visualization is composed of multiple individual trend streams, or *trendlines*, represented by sinuous and variable-width lines that branch and

- *W. Javed is with General Electric, San Ramon, CA, USA. E-mail: javed@ge.com.*
- *N. Elmqvist is with University of Maryland, College Park, MD, USA. E-mail: elm@umd.edu.*
- *S. Ghani is with the KACST GIS Technology Innovation Center, Umm Al-Qura University, Makkah, Saudi Arabia. E-mail: sghani@gistic.org.*
- *K. N. Hampton is with the Rutgers University, New Brunswick, NJ, USA. E-mail: keith.hampton@rutgers.edu.*
- *S. Gad, T. Ewing, and N. Ramakrishnan are with Virginia Tech, Blacksburg, VA, USA. E-mail: {samah, etewing, naren}@vt.edu.*

*Manuscript received 2 Jan. 2014; revised 26 Nov. 2014; accepted 21 Dec. 2014. Date of publication 31 Dec. 2014; date of current version 1 Apr. 2015. Recommended for acceptance by H. Qu.*
*For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.*
*Digital Object Identifier no. 10.1109/TVCG.2014.2388208*

merge in a fashion reminiscent of real-world river deltas. We use the width of each trendline to communicate the prominence of its trend in the dataset at a particular time, and its color to communicate category or overall weight of the trend. A heuristic layout technique calculates the order of trends for each timestamp while minimizing the number of trendline crossings. Interaction techniques allow for highlighting individual trendlines, changing the layout order, and drilling down into the data.

As demonstrated later, there are several potential usage scenarios for our novel ThemeDelta system. We dedicate the main part of this paper to explore three scenarios: (1) historical U.S. newspaper data from four months in the year 1918 during the second wave of the Spanish flu pandemic; (2) the similarities and differences in trends and themes being discussed by the two candidates in the U.S. 2012 presidential campaign; and (3) social messages exchanged between virtual communities via the i-Neighbors web-based application [4]. These applications are intended to demonstrate that ThemeDelta provides an interesting insight into datasets not immediately apparent through other representations.

The remainder of this paper is structured as follows: We first survey the literature on text visualization, temporal visualization, and temporal topic modeling. We then present an overview of ThemeDelta system followed by detailed review of the temporal topic modeling algorithm and visual representation. The next three sections deal with the presidential campaign, iNeighbors, and newspaper applications. We then discuss the results of a qualitative user study, performed to measure the effectiveness of the system when used by experts. We close with a discussion, conclusions, and future plans.

## 2 Related Work

Textual data is one of the most common forms of data today [5] and is ubiquitously available from a wide variety of sources such as websites, books, reports, newspapers, and magazines. Reading and understanding many textual datasets is often impossible due to their massive scale or having limited time to review them [6]. Tools such as information retrieval, data mining, and visualization have been suggested to help in this activity, often by summarizing texts and highlighting important themes. Here we survey the most relevant among such tools.

### 2.1 Text Visualization

Text visualization uses interactive visual representations of text that go beyond merely the words themselves to show important features of a large dataset [7]. The simplest feature is word frequency; tag clouds (or word clouds) [8], [9] are the most common text visualization technique and use graphical variables like font, color, weight, and the size of the word to convey its importance in the source text. Variations include Wordle [10] and ManiWordle [11] for producing more compact and aesthetically pleasing layouts. More advanced text visualization techniques convey not only word frequency but also structural features of the text corpus (or corpora), often using a graph structure. Examples include [12], Phrase Nets [13], and GreenArrow [14], which all used variants of a dynamic graph to convey the structure in the document. Clustered word clouds [15], [16] took another approach by using word structure to modify the layout in a tag cloud, which is similar to the themescape representation used by IN-SPIRE [17].

### 2.2 Temporal Text Visualization

Text data may often be analyzed over time to expose aspects of the data that evolved during a specific time period; for example, recent studies have highlighted the importance of time and causality in investigative analysis [18]. While less saturated than general text visualization, several pieces of work exist that visualize text over time in this field.

Perhaps one of the earliest and most seminal of these is ThemeRiver [1], which essentially can be thought of as a horizontal arrangement of a large number of one-dimensional tag clouds over time; the keywords become bands in a stacked graph evolving on a timeline. Byron and Wattenberg [2] later formalized these representations into so-called streamgraphs, and applied them to many other types of data. Several approaches exist that are similar to the basic ThemeRiver stacked chart. Tufte [19] showed an illustration of changing themes of music through the ages. NewsLab [20] applied a ThemeRiver representation to visualize a collection of thousands of news videos over time. NameVoyager [21] used streamgraphs to show temporal frequencies of baby names. TIARA [22] blended tag clouds onto temporal stacked charts for important themes. However, because stacked charts group themes into a single visual entity, none of these techniques are capable of conveying the structural features of the corpus.

A select few techniques provide both temporal and structural information. Parallel tag clouds (PTCs) [23] integrated one-dimensional tag cloud layouts on a set of parallel axes, each which could potentially be used for a time or date. However, while visually similar to our ThemeDelta technique, PTCs did not explicitly convey the grouping of words into topics. Another similar design is TimeNets [24], which used colored lines to show groupings over time for genealogical data. In fact, "Movie Narrative Charts (comic 657) of the web comic XKCD[1] also used sinuous lines to convey groupings of characters in time and space for famous movies. Finally, Turkay et al. [25] presented two techniques for visualizing structural changes of temporal clusters that are reminiscent of ThemeDelta; while not specifically designed for text, the clusters used in their work could very well stem from textual corpora. ThemeDelta uses a similar visual metaphor but focuses on text and is intrinsically tied to the scatter/gather temporal segmentation component as well.

A very relevant work is ParallelTopics, a visual analytics system that integrated LDA with interactive visualization [26]. The system used the parallel coordinate metaphor to present document topic distributions, with applications to exploring National Science Foundation awarded proposals, VAST publications, as well as tweets. This system presented the underlying probabilistic distributions in the LDA model from a temporal perspective using multiple aggregation strategies and interactions. The system can capture the topics and their evolution over time but only by using fixed

1. Available online at: http://xkcd.com/

time frames. In contrast, our ThemeDelta approach discovers time frames automatically based on topic reorganizations across time.

Finally, TextFlow [27] is perhaps the closest related work to ThemeDelta and used tightly integrated visualization and topic mining algorithms to show an evolving text corpus over time. However, whereas we draw upon the same basic visual representation as TextFlow, our focus in this work is segmenting time based on topic shifts and then interfacing with standard topic modeling using a novel algorithm. Furthermore, ThemeDelta does not aggregate keywords into stacks or glyphs but instead puts more emphasis on an interactive layout.

### 2.3 Temporal Topic Modeling

Text visualization techniques are only as good as the extracted features they are visualizing. *Text mining* is the concept of deriving high-quality features from text [28]. One of the current most promising lines of research in text mining is *topic modeling* [3], where documents are modeled as distributions (mixtures) over topics, and topics in turn are distributions over the vocabulary used in the corpus. LDA is considered a generalization of Probabilistic Latent Semantic Analysis (PLSA), proposed by Hofmann [29]. The difference between LDA and PLSA is that the topics distributions in LDA are assumed to be distributed according to a Dirichlet prior.

Temporal topic modeling algorithms started to appear around 2006, with most of them being generalizations of static topic models. The difference between our goals and those of previous work is that we aim to automatically identify segment boundaries that denote shifts of coverage and, in this manner, extract temporal relationships for examination. Therefore, we are not proposing a new topic model but instead are proposing how we can "wrap around existing topic modeling algorithms to segment timestamped data.

Classical work in this space was done by Blei and Lafferty [30], who extended traditional state space models to identify a statistical model of topic evolution. They also developed techniques for approximating the posterior inference for detecting topic evolution in a document collection. Wang and MacCallum [31] proposed a non-Markov model for detecting topic evolution over time. They assume that topics are associated with a continuous distribution over timestamps and that the mixture distribution over topics that represent documents is influenced by both word co-occurrence relationships and the document timestamp. Thus, in their model, topics generate both observed timestamps and words. Iwata and Yamada [32] also proposed a topic model that enabled sequential analysis of the dynamics of multiple time scale topics. In their proposed model, topic distributions over words are assumed to be generated based on estimates of multiple timescale word distributions of the previous time period. Finally, Wang et al. [33] have recently proposed a model that replaced the discrete state space that was originally proposed by Blei and Lafferty [30] with a Brownian motion law [34] to model topic evolution. They assumed that topics are divided into sequential groups so that topics in each slice are assumed to evolve from the previous slice.

This line of research has been extended to mining text streams, e.g., as done in [35]. Here, the authors study the problem of mining evolving multi-branch topic trees inside a text stream by proposing an evolutionary multi-branch tree clustering method. In their method, they adopt Bayesian rose trees to build multi-branch trees and used conditionals prior over tree structures to keep the information from previous trees as well as maintain tree smoothness over time. To keep the consistency of trees over time, they define a constraint tree from triples and fans to compute the tree structure differences.

Other work focused on multiple text streams. Wang et al. [36] and [37] aimed to "align time series streams so as to identify correlated and/or common topics across disparate streams. Our algorithm was designed to analyze a single time-indexed corpus as opposed to multiple time (and asynchronous) text streams. The work of Leskovec et al. [38] stringed together individual tweets (meme) into a thread. The authors placed additional constraints in identifying these threads (e.g., that they originate in a single meme). The granularity of analysis in our work is clusters rather than tweets; the desired output is rich scatter-gather relations between clusters rather than simple branching patterns. The work presented by Gao et al. [39] is the closest to our work. The authors conducted topic modeling and investigate both scattering and gathering possibilities of cluster organization. The difference is that our work automatically determines segmentation boundaries where significant shifts of topic distributions occur, whereas the work of Gao et al. incrementally clustered every time point separately and then aimed to make splitting and merging decisions.

## 3 ThemeDelta Overview

ThemeDelta is intended to convey local and global temporal changes in the distribution of evolving trends. The system detects and visualizes how different trends converge and diverge into groupings at different points in time, as well as how they appear and disappear during a time period. The system consists of two major components: a backend data analytics component, and a frontend visualization component:

- The *analytics backend* is responsible for accepting a large temporal text corpus and automatically identifying segments that characterize significant shifts of coverage. The algorithm responsible for this task was originally developed to detect deliberation in social messages [40]. For ThemeDelta it was modified to detect and work with named entities: geopolitical, person, organization, as well as possibly uncategorized entities.
- The *visualization frontend* is responsible for graphically representing the discovered trending of topics. While originally designed for timestamped text collections, we see many additional applications such as for genealogy (e.g., [24]), communication graphs (e.g., [41]), and general dynamic graphs.

In this section, we review the data format expected by the different components of the system and then discuss implementation details.

### 3.1 Data Format

The backend accepts a text dataset consisting of timestamped data. The frontend takes the output of the backend for visualization. The backend output consists of *trends* , *topics* (groups of trends at a specific point in time), and *segments* (a closed interval of time, modeled as a group of topics).

The exact mapping of these general concepts to a dataset is domain-specific. For example, for a timestamped document collection, trends could represent the terms extracted from the documents, and the topics would model how these terms converge into groupings at different points in time.

### 3.2 Implementation

ThemeDelta is web-based application that is built to be capable of running in any modern web browser. To realize this the backend of ThemeDelta was built using Java and this includes all data preprocessing, clustering, and segmentation.

The frontend was built using JavaScript and SVG. The current implementation is fully interactive and animated, and is built using the Raphaël[2] toolkit for scalable vector graphics.

## 4 Temporal Topic Modeling

The ThemeDelta backend is responsible for the segmentation and discovering the trending topics from the input dataset.

Our segmentation algorithm expects the input data to be in a bag-of-word format. The preprocessing needed is thus to tokenize the text into individual words, followed by standard processing steps such as: lower case conversion, stemming, stop words removal, spell checking, and punctuation removal.

The main task of the segmentation algorithm is to automatically partition the total time period defined by the documents in the collection such that segment boundaries indicate important periods of temporal evolution and reorganization.

The algorithm moves across the data by time and evaluates two adjacent windows assuming a given segmentation granularity (e.g., discrete days, weeks, or months). This granularity varies from application to another and it is decided by domain experts. We evaluate adjacent windows by comparing their underlying topic distributions and quantifying common terms and their probabilities.

We chose to quantify common terms based on the overlap between them. The overlap can be captured using a contingency table. Fig. 2 shows a simplified example of two segments, each comprising three topics and the corresponding contingency table measuring the overlap between these distributions. For example, topic 1 $(Z1)$ in segment 1 and topic 1 $(Z'1)$ in segment 2 overlap in $w1$ and $w6$. This resulted in adding the count 2 in the contingency table cell that corresponds to the overlapping cell between the two topics from the two segments. We would like the topic models of the two adjacent windows to be maximally independent, which will happen if the table entries are near uniform.

Formally, given the input data to be indexed over a time series $T = \{t_1, t_2, \ldots, t_t\}$, the segmentation problem we are trying to tackle is to express $T$ as a sequence of segments or windows: $S_T = (S^{t_a}_{t_1}, S^{t_b}_{t_{a+1}}, \ldots, S^{t_l}_{t_k})$ where each of the windows $S^{t_e}_{t_s}, t_s \leq t_e$ denotes a contiguous sequence of time points with $t_s$ as the beginning time point and $t_e$ as the ending time point.

Each window $S^{t_e}_{t_s}$ has a set of topics that is discovered from the set of documents that fall within this window. The topics are discovered by applying LDA [3]. Applying this algorithm will result in two main distributions: document-topic distribution (representing the distribution of the discovered topics over the documents) and topic-terms distribution (representing the distribution of the discovered topics over the vocabulary).

Topics within each window is represented as $S^{t_e}_{t_s} = \{z_1, z_2, \ldots, z_n\}$ where $n$ is the number of top topics $z$ discovered. Each topic is represented by a set of terms $w$ as follow: $z_i = \{w_1, w_2, \ldots, w_m\}$ where $m$ is the number the top terms extracted from the topic-terms distribution resulted from applying LDA on the documents within a window. Number of top topics $n$ and top terms representing a topic $m$ vary from application to another.

We represent two adjacent windows as $S^{t_{e_1}}_{t_{s_1}}$ and $S^{t_{e_2}}_{t_{s_2}}$. To evaluate two adjacent windows, we construct the contingency table for two windows. The contingency table is of size $r \times c$ where rows $r$ denote topics in one window and columns $c$ denote topics in the other window. Entry $n_{ij}$ in cell $(i, j)$ of the table represents the overlap of terms between topic *i* of $S^{t_{e_1}}_{t_{s_1}}$ and topic *j* of $S^{t_{e_2}}_{t_{s_2}}$.

We used a contingency table because it enable the replacement of LDA with any emerging topic modeling variants. As presented in [42] we can embed any vector quantization clustering algorithm in a contingency table framework. For instance, distributions inferred from a more sophisticated model can be compared using the contingency table formulation introduced here.

Then to check the uniformity of the table, three steps should be accomplished: first, calculate the following two quantities:

- Column-wise sums $n_{i.} = \sum_j n_{ij}$
- Row-wise sums $n_{.j} = \sum_i n_{ij}$

These two quantities will be used to quantify the overlap between the topics discovered from two adjacent windows. In our implementation for this step, each topic is represented by its top assigned terms. The contingency table is created from these terms (here we chose 20 terms and the choice of the number of terms is inherently heuristic and specific to the application). A probabilistic similarity measure such as the KL- or JS-divergence between the distributions being compared is another possibility.

Second, we define two probability distributions, one for each row and one for each column:

$$p(R_i = i) = \frac{n_{ij}}{n_{i.}}, (1 \leq j \leq c) \tag{1}$$

$$p(C_j = j) = \frac{n_{ij}}{n_{.j}}, (1 \leq i \leq r). \tag{2}$$

2. Available online at: http://raphaeljs.com/

Third, we calculate the objective function $F$ to capture the deviation of these row-wise and column-wise distributions with regard to the uniform distribution.

The objective function is defined as follows:

$$F = \frac{1}{r}\sum_{i=1}^{r} D_{KL}\left(R_i \| U\left(\frac{1}{c}\right)\right) + \frac{1}{c}\sum_{j=1}^{c} D_{KL}\left(C_j \| U\left(\frac{1}{r}\right)\right) \quad (3)$$

where

$$D_{KL}(P\|Q)) = \sum_i P(i)\log\frac{Q(i)}{P(i)}. \quad (4)$$

This objective function can reach a local minimum, which is acceptable given that we are trying to segment time based on shifts in topics and this approach capture the first shift in topics (as opposed to detecting an optimal segmentation which would require a more exhaustive search through breakpoint layouts).

Here, $D_{KL}$ denotes the KL-divergence that is used to calculate the distance between the row-wise and the uniform distribution. Likewise, it is used to calculate the distance between the column-wise distributions and the uniform distribution. Then the values resulting from using the $D_{KL}$ will be used in calculating the objective function $F$.

The algorithm repeats the above mentioned steps for all permutations of the two sliding window sizes. The goal is to minimize $F$, in which case the distributions observed in the contingency table are as close to a uniform distribution as possible, in turn implying that the topics are maximally dissimilar.

There are two stopping conditions for this algorithm: (1) if conversion of $F$ is achieved, or (2) the maximum size for both windows was achieved. Detailed description of the algorithm is shown in Algorithm 1.

Putting all these ideas together into a working ThemeDelta application involves several implementation decisions. Model selection (i.e., choosing the 'right' number of topics) is a key underlying issue and Bayesian criteria such as the AIC or BIC or information-theoretic criteria such as perplexity can be utilized for this purpose. More formally, a hierarchical Dirichlet process (HDP) model can be employed. An inordinate number of topics would lead to an under-determined system and inference of redundant topics. Here, we apply a simple word filtering technique on the discovered topics within and across segments to eliminate repeated topics and repeated terms. This can potentially alter the number of topics in each segment and number of words used to represent each topic.

## 5 ThemeDelta: Visual Representation

ThemeDelta's visual representation draws on TextFlow [27], and uses a basic visual encoding consisting of sinuous *trendlines*—each representing a trend in the dataset—stretching from left to right along a timeline mapped to the horizontal axis (Fig. 3). The horizontal space along this axis is divided equally among different time segments ($t_1$, $t_2$, and $t_3$ in Fig. 3). Topics for each segment are perceptually conveyed by clustering the trendlines for the grouped trends next to each other along the vertical axes, leaving a fixed amount of empty vertical space between adjacent topics. Vertical lines, one for each time segment, partition their horizontal positions.

**Algorithm 1.** Topic Modeling Based Segmentation

**Input:** $T = \{t_0, t_1, t_2, t_3, \ldots, t_t\}$
$x =$ min. window size.
$y =$ max. window size.
**Output:** $S_T = \{\}$ //Set of all segments between $t_0$ and $t_t$
$W1Start = t_0$
$W1Size = x$
$F =$ Initialize objective function with a large number.
**while** $W1Start + W1Size + x \leq t_t$ **and** $W1Size \leq y$ **do**
//$x$ is added to W1 to take into account the data availability for W2.
  Conversion = False
  //Reset start and size of W2.
  $W2Start = W1Start + W1Size + 1day$
  $W2Size = x$
  **while** $W2Start + W2Size \leq t_t$ **and** $W2Size \leq y$ **do**
    Apply LDA separately on $W1$ and $W2$
    Calculate $F'$ for $W1$ and $W2$
    **if** $F' > F$ **or** $W1Size == y$ **or** $W2Size == y$ **do**
    //Conversion or max. window size limit reach.
      Add $W1$ and $W2$ to $S_T$
      $W1Start = W2Start + W2Size + 1day$
      $W1Size = x$
      Conversion = True
      **Break**
    $F = F'$
    $W2Size+ = x$ //Expand W2.
  **if** $!Conversion$ **do**
    $W1Size+ = x$
 **if** leftover data exists **do**
  //leftover data starts at W1Start and ends at $t_t$.
  Apply LDA on leftover data.
  Add window of leftover data to $S_T$.
 **return** $S_T$

### 5.1 Visual Design

Given this basic design, many design parameters remain open. Below we review the most important of these and motivate our decisions for the visualization technique. A visualization developer using the same basic visual representation may make different choices than these depending on the application.

*Shape.* To communicate the organic nature of evolving trends, we use splines to yield smooth curves. The resulting lines are continuous, predictable, and appealing. An alternative design would have used rectilinear or sharp angles, but curves are likely easier to perceive and more aesthetically pleasing.

*Thickness.* Trendline thickness is a free visual variable. While it is possible to use a uniform thickness for all trendlines, it can also be used to convey scalar data for each time segment. Because increasing thickness will raise the visual salience of a trendline, we tend to use it to convey the weight of each keyword calculated by our segmentation algorithm.

Furthermore, our visual representation uses vertical dashed lines to partition time segments on the visual space.

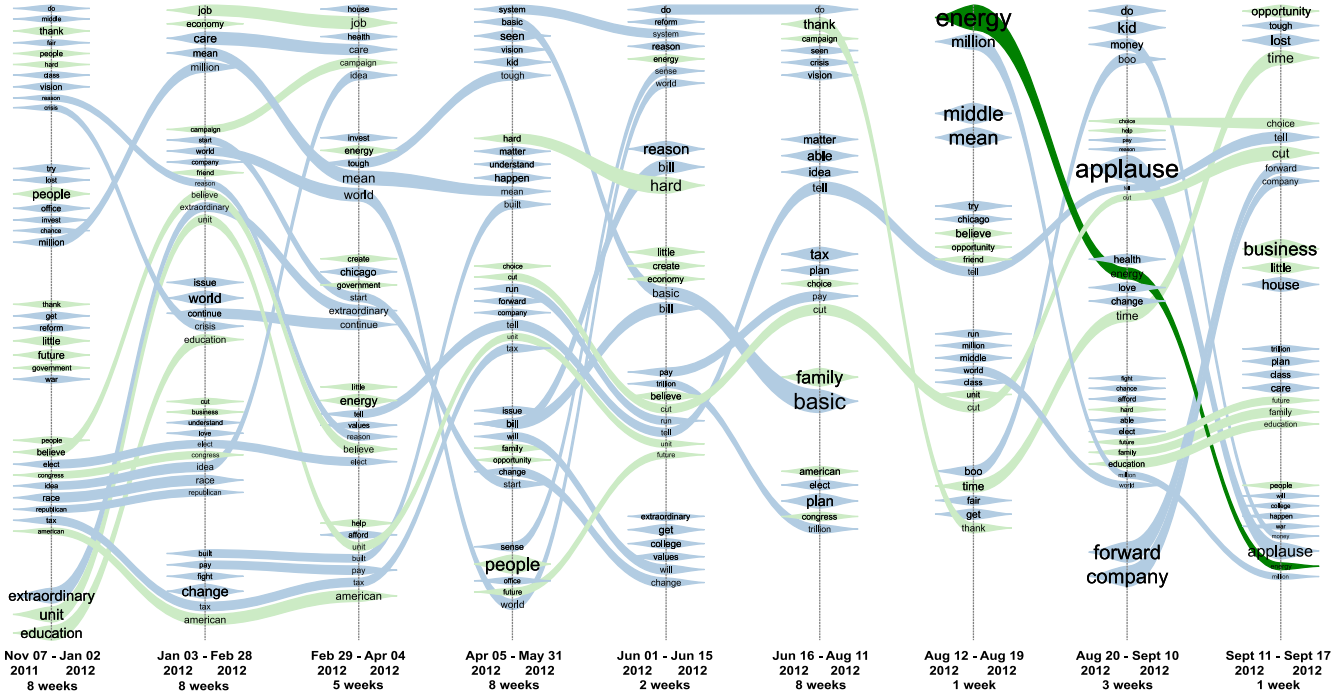



Fig. 1. ThemeDelta visualization for Barack Obama' campaign speeches during the U.S. 2012 presidential election (until September 10, 2012). Green lines are shared terms between Obama and Romney. Data from the "The American Presidency Project" at UCSB (http://www.presidency.ucsb.edu/).

The thickness of these lines is another free variable that can, e.g., be used to indicate the relative extent of each time segment. This is useful since time segments may be irregular; some segments are significantly longer than others.

*Color.* Color is another free parameter in our visual representation, and can convey either a quantity (using a color scale) or a category (using discrete colors). The choice depends on the application. For example, we use it both to show the strength of a correlation, as well as to convey which entity class a particular trendline belongs to.

*Discontinuities.* Trendlines can begin and end at any time segment, sometimes only to reappear later in time. We communicate this using a tapered endpoint of the line (see borders in Fig. 3). An alternative design could have dashed the trendlines for the periods of time where there is no associated value, similar to the use of different trend shapes in TextFlow [27]. We chose to avoid this to minimize visual complexity.

*Labels.* We draw the names of each trendline on the line itself for each time segment. While this is redundant (one instance of the label is sufficient) and potentially a source of visual clutter, it prevents the user from having to trace an undulating trendline back to a single label at the far end of the visualization. We also scale the label size based on the trendline's thickness, similar to word scaling in word clouds.

*Duplicated Trends.* Sometimes a trend may exist in more than one topic for a particular time segment (see trend $A$ at time $t_2$ in Fig. 3). To make the visualization consistent, as well as to convey the fan-out, we are forced to fork the trendline into two or more pieces. Analogously, in a time segment following a duplicated trend instance, the trendlines should be merged to maintain consistency. In situations when there is more than one candidate to fork from or merge to, we choose the two trend instances that are vertically closest to each other (see the layout algorithm discussed below).

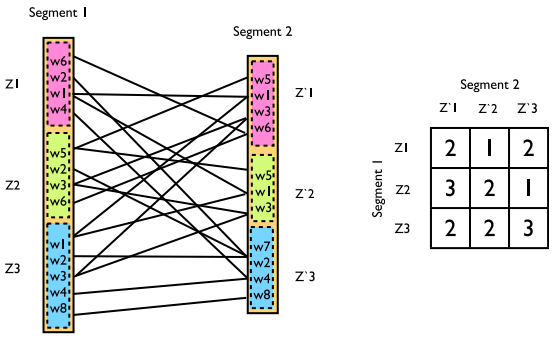



Fig. 2. Contingency table used to evaluate independence of topic distributions for two adjacent windows [40].

## 5.2 Interaction

Several interaction techniques are meaningful for ThemeDelta frontend. First of all, geometric zoom and pan allows for being able to magnify a certain part of the visualization to see details. Furthermore, hovering over a trendline will highlight the line, including all of its branches in other parts of the visualization (even past a discontinuity). Fig. 1 shows this interaction, where the trend lines associated with the keyword *energy* are highlighted in response to a mouse hover interaction.

The interface also supports searching for trendlines by name. In addition, we provide a combined filtering and resorting operation. Clicking on a trendline will add it to a filter box, causing the layout to be recomputed with the selected trendline at the top of the screen. The new layout will only include trendlines that are connected to the selected trendline, i.e., which in at least one time segment

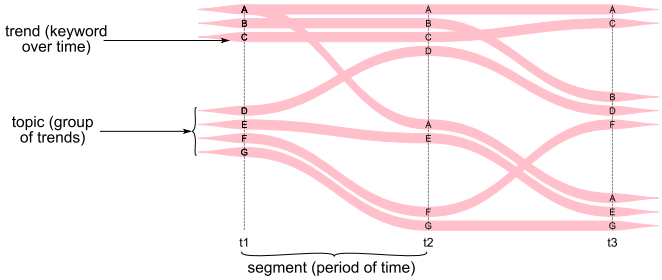



Fig. 3. Basic visual representation used by ThemeDelta.

belong to the same topic as the selected trendline. Following the example presented in Fig. 1, clicking on the trend line for the keyword *energy* performs the filtering operation, and the visual layout is changed such that the filter keyword is positioned at the top (Fig. 4). Additional trendlines can be added to the filter box, yielding a conjunctive filter (only trendlines which are connected to *all* selected trendlines are shown).

### 5.3 Layout

The ThemeDelta frontend visualization layout divides the available horizontal space equally between time segments, while vertical space is divided locally between the topics associated with each time segment. Due to the ever-changing topic groupings over time as well as the dynamic appearance and disappearance of trends, it is typically not possible to represent trends as straight lines. In fact, a single trend could appear in a different topic and at a different vertical position with each new time segment. This is the reason for using smooth splines to convey this organic trend evolution.

Of course, this in turn means that trendlines will frequently cross one another while connecting the multiple occurrences of a single term across different time segments. Research in graph drawing has shown that the ease with which a user can follow an edge depends on the number of crossings with other lines in its path [43].

Tanahashi and Ma [44] discussed a set of layout design principles for better legibility of storyline visualizations like ThemeDelta. However, the complexity of their algorithm makes it difficult to achieve real-time layout updates. Other work proposed by Liu et al. [45] trade-off optimal layout with algorithm performance to achieve real-time updates. The algorithm used for ThemDelta is similar to the one proposed by Liu et al. [45]. However, contrary to their algorithm, we do not have hierarchical relationships in the underlying data and to facilitate the identification of individual topics by supporting a constant reasonable vertical space between them, the layout algorithm used in this paper does not perform the topic alignment step. Moreover, to achieve real time interactivity our implementation minimizes line crossings through a single iteration across different time segments.

In particular, ThemeDelta relies on a deterministic layout algorithm that minimizes trendline crossings by first sorting the vertical positioning of different topics, followed by sorting the trends within each topic. While sorting topics at a particular time segment $t_i$, a topic $p_1$ is placed before another topic $p_2$ if the average vertical position of the trends contained in topic $p_1$ is less than the terms present in topic $p_2$ at the previous time segment $t_{i-1}$. Topic position in the first time segment is either determined randomly, or using some attribute of the underlying data.

After sorting topics it is time to sort the trends within each topic. Except for the first time segment, trends within a topic are sorted such that their relative vertical position remains the same as it was in the previous time segment. Once all trends are sorted, the trends contained within topics of the first time segment are sorted such that their vertical position remains the same as in the second time segment.

Fig. 5 shows the progressive decrease in the number of trendline crossings at different stages of the layout. In Fig. 5a the dataset is visualized without any sorting. This results in a total of twelve crossings between trendlines, connecting multiple occurrences of terms across the two time segments $t_1$ and $t_2$. Fig. 5b shows the resulting layout after topic sorting. As shown in the figure, the topics within time segment $t_2$ are now ordered based on the average vertical position of their corresponding terms within time segment $t_1$. This ordering of topics has reduced the number of line crossings from 12 to 6. Finally, Fig. 5c shows the resulting layout after trend sorting. Here again it is evident that the number of line crossings is reduced even further. All in all, as a result of the layout algorithm, the number of trendline crossings has been reduced from 12 to 2.

## 6 U.S. 2012 Presidential Campaign

Political speeches, especially during an election campaign, are particularly interesting document collections to

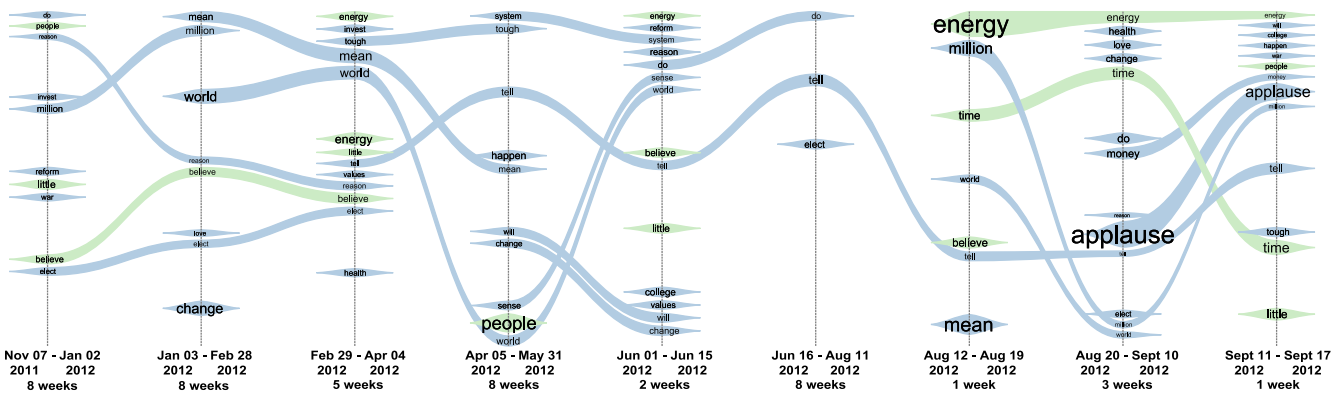
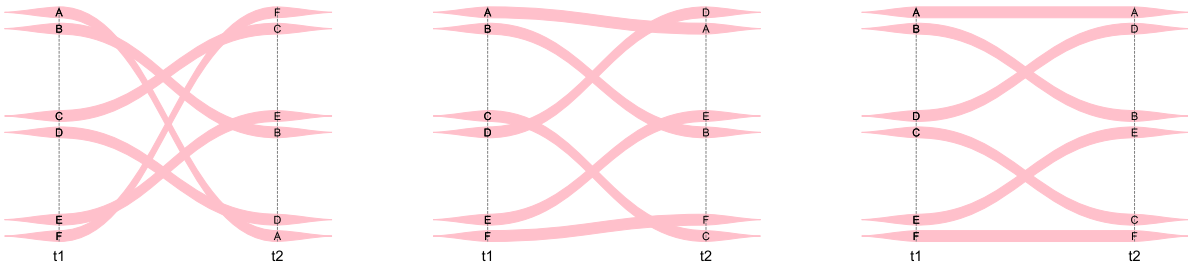



Fig. 4. ThemeDelta visualization after performing a filtering operation, based on the keyword “energy”, in the visualization presented in 1.




(a) Without sorting. (b) Topic sorting only. (c) Topic plus term sorting.

Fig. 5. Comparison of different stages of the layout sorting algorithm used for the ThemeDelta technique.

analyze because the political discourse tends to change and evolve as different candidates respond and challenge each other over the course of the campaign. Visualizing the speeches of different candidates would allow for comparing the trends of each candidate with each other. To study such effects, we used the U.S. 2012 presidential election campaign speeches.

The U.S. presidential election takes place every four years (starting in 1792) in November (the 2012 election day was November 6), and is an indirect vote on members of the U.S. Electoral College, who then directly elect the president and vice president. In 2012, the Republican and Democratic (the two dominant parties, representing conservative vs. liberal agendas) conventions were held on the weeks of August 27 and September 3, respectively. The two opposing candidates were Republican nominee Mitt Romney, and Democratic nominee Barack Obama (incumbent President of the United States). The ThemeDelta for both candidates is shown in Figs. 1 and 6.

### 6.1 Data

In collecting data for the United States presidential election, we used campaign speech transcripts for both candidates from the UCSB American Presidency Project.[3] For Mitt Romney, we used transcripts from 46 speeches over a 62-week period: from announcing candidacy on July 29, 2011, to August 14, 2012. This corpus included speeches from both the Republican primary election (settled on May 14, 2012 as the main competing nominee Ron Paul withdrew). For Barack Obama, we used transcripts from 40 speeches over a 44-week period: November 7, 2011 to September 17, 2012.

### 6.2 Discussion

Visualizations of the two candidates Barack Obama and Mitt Romney are shown in Fig. 1 and Fig. 6. Trendlines in both visualizations represent characteristic keywords that each candidate uses as a theme in his speeches. Democratic trendlines are colored blue, Republican ones are red, and trendlines for keywords that both candidates share are green.

For the Romney dataset (Fig. 6), there is a clear impact of time on keywords and topics that the candidate is using. Romney's message starts out relatively simple with only two main topics, but quickly branches out in complexity as time evolves. The effect of main competitor Ron Paul withdrawing in May is clear: before this date, Romney is trying to win the party nomination, whereas afterwards, he is going for the presidential seat. As a result, his message becomes more simple again: both the number of keywords and the number of topics decreases during the last three segments, presumably to focus on key issues in the Republican election platform.

For the Obama dataset (Fig. 1), a good portion of the identified keywords are common with Mitt Romney (i.e., green in color). This could be seen as Obama discussing many of the issues that has become central to the U.S. presidential race. Furthermore, there is a clear presence of keywords such as "health," "insurance," and "care," which may refer to the president's health care reform from 2010 (informally called Obamacare). This is a controversial issue that still causes a major divide between voters; a Reuters-Ipsos poll in June 2012 indicated that a full 56 percent of Americans were against the law.

Taken as a whole, both datasets have a heavy emphasis on economics keywords. This is commensurate with the overall theme of the 2012 presidential race, which largely has focused on the poor economic situation of the United States.

## 7 i-Neighbors Social Messages

The Internet facilitates informal deliberation as well as civic and civil engagement. Web-based applications for informal deliberation (e.g., i-Neighbors [4]) facilitate the collection of data that we can analyze to provide insight into how neighborhoods with different poverty levels use ICTs for informal deliberation. Using ThemeDelta, we can characterize differences and detect common interests in informal deliberation between advantaged and disadvantaged communities.

The goal of this application was to study two basic questions: what lengths of time neighborhoods with different poverty levels spend discussing topics? And what is the average similarity in topics discussed between neighborhoods with different poverty levels, and the similarity in topics discussed between neighborhoods with similar poverty levels?

### 7.1 Data

The data for this application was collected through the i-Neighbors system. The site is open for anyone in the United States or Canada to create a virtual community that matches

3. Available online at: http://presidency.ucsb.edu/

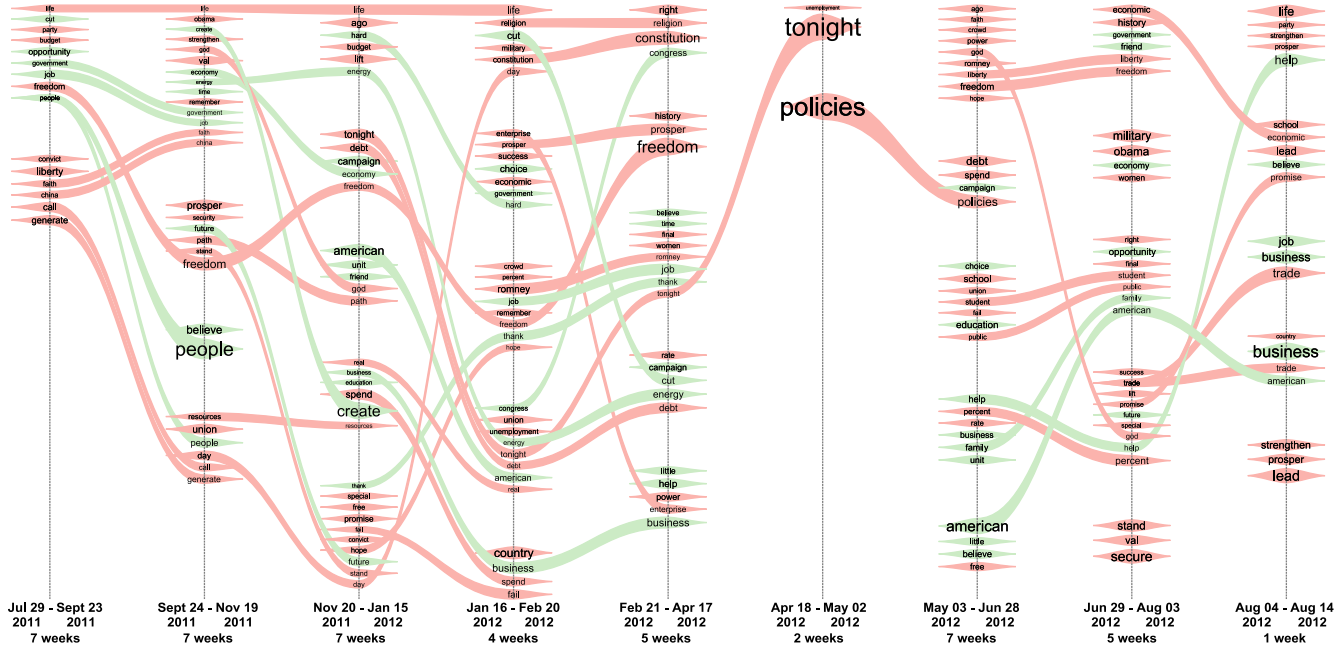



Fig. 6. ThemeDelta visualization for Mitt Romney campaign speeches for the U.S. 2012 presidential election (as of September 10, 2012). Green lines are shared terms between Obama and Romney speeches. Data from the American Presidency Project at UCSB (http://presidency.ucsb.edu/).

their geographic neighborhood and start conversations on local issues, announce events, disseminate information, etc.

When we collected the data in 2010, the i-Neighbors website had over 100,000 users who had registered more than 15,000 neighborhoods. Over 1,000 neighborhoods were active with more than 7,000 unique messages contributing to neighborhood discussion forums. We collected data from six geographically diverse communities located in Georgia, Maryland, New York, and Ohio. We selected the three groups located in areas with concentrated levels of poverty (a poverty rate of 25 percent or more, 2009 American Community Survey, US Census Beureau) who exchanged the most messages, and the three most active groups in more advantaged areas.

### 7.2 Discussion

We applied our temporal segmentation algorithm on the six selected neighborhoods. Topics within each segments can be examined using the visualization to find topic similarities between neighborhoods. Segmentation labels indicating segments size can be used for comparing the time spent by different neighborhood discussing certain topics.

A partial segmentation output is shown for a disadvantaged neighborhood in Figs. 8 and 9 for a more advantaged neighborhood. From these two examples, the segments sizes are not very different and we can conclude that both the disadvantaged and advantaged neighborhoods spend similar amounts of time discussing topics.

Examining the words groupings in both neighborhoods can lead to discovering differences and similarities in their discussions. For example, in the low-poverty neighborhood in segment [Feb 1, 2009 to Nov 1, 2009], there is a topic that has the words "watch" and "neighbor," which lead us to conclude that there were some arrangements or discussions about a neighborhood watch. This topic is not found in the the disadvantaged neighborhood visualization. If the user searched for the word "watch" this will result (Fig. 7) in only showing the topics that has the this word and any other related topic.

Similarly, an example of similarities of topics discussed between neighborhoods can be shown by examining the segment [Jan 03, 2010 to Sept 3, 2010] in the advantaged neighborhood (Fig. 9) and segment [Feb 4, 2010 to Oct 4, 2010] in the disadvantaged neighborhood (Fig. 8). In both segments, there exist two topics in which both communities discuss a park-related project.

## 8 Historical U.S. Newspapers

Newspaper stories are precisely the type of ongoing, evolving trend datasets for which ThemeDelta was designed. Below we review the source, segmentation, and visualization for a dataset consisting of historical U.S. newspaper stories from 1918.

### 8.1 Data

Our data source was a historical newspapers database hosted on the Chronicling America website,[4] a Library of Congress resource that provides an Internet-based, searchable database of historical U.S. newspapers. The website is maintained by the National Digital Newspaper Program (NDNP), a partnership between the National Endowment for the Humanities (NEH) and the Library of Congress.

Some of newspapers included in this example are: *The Washington Times* (Washington, DC), *Evening Public Ledger* (Philadelphia, PA), *The Evening Missourian* (Columbia, MO), *El Paso Herald* (El Paso, TX), and *The Holt County Sentinel* (Oregon, MO). We gathered data from them, restricting the time to the period September 1918 through December 1918. From this dataset, we extracted only paragraphs that mention the word "influenza" resulting in 2,944 paragraphs. This corresponds to the 1918 flu pandemic (also known as the "Spanish flu")

4. Available online at: http://chroniclingamerica.loc.gov/

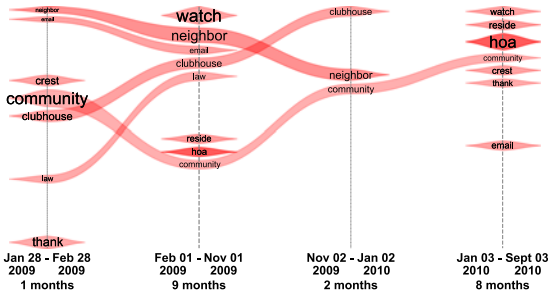



Fig. 7. Result of searching for the word "watch" in low-poverty neighborhood.

which spread around the world from January 1918 to December 1920, resulting in some 50 million deaths.

## 8.2 Discussion

Applying the above dataset to ThemeDelta using a weekly segment granularity yields four discrete time segments over the four-month time period. Fig. 10 shows a visualization of the result, where the transparency value of each trendline has been mapped to the global ranking of the keyword corresponding to the trendline. The thickness of the trendline conveys the ranking of each keyword for a particular time segment, calculated by our segmentation algorithm.

Fig. 10 offers several observations that summarize the qualitative nature of trends exposed by ThemeDelta. The output is showing many events that were related to the 1918 pandemic in the data. For example, in the first time segment, September 9 until October 9, there are a topic that contain the terms "mask" and "German." This corresponds to advisories and guidelines recommending people to use masks to protect themselves from the ongoing influenza pandemic during World War I. In the same segment, the words "liberty," "loan," and "campaign" appeared in one

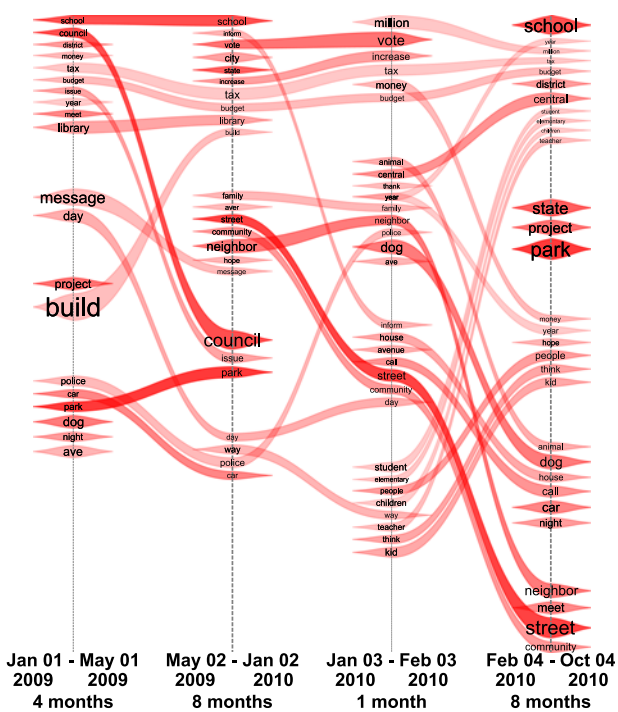



Fig. 8. Partial output from a high-poverty neighborhood.

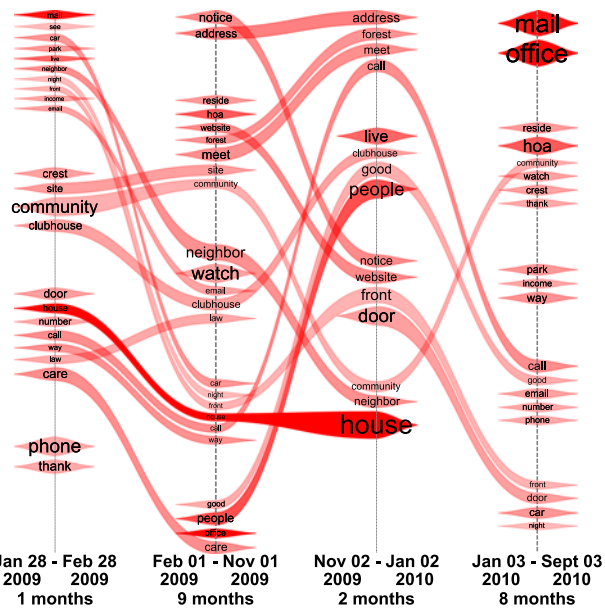



Fig. 9. Partial output from a low-poverty neighborhood.

of the topics, and continued appearing in the following segment, October 10 until December 5, because a liberty loan campaign were issued to support the army during World War I. Also, in the October 10-December 5 segment, the army men left the camps to go back home from service and stay with their families; this explains the topic with the words "family," "home," "serves," "spent," "wife," and "son." This topic appeared along with the topic with

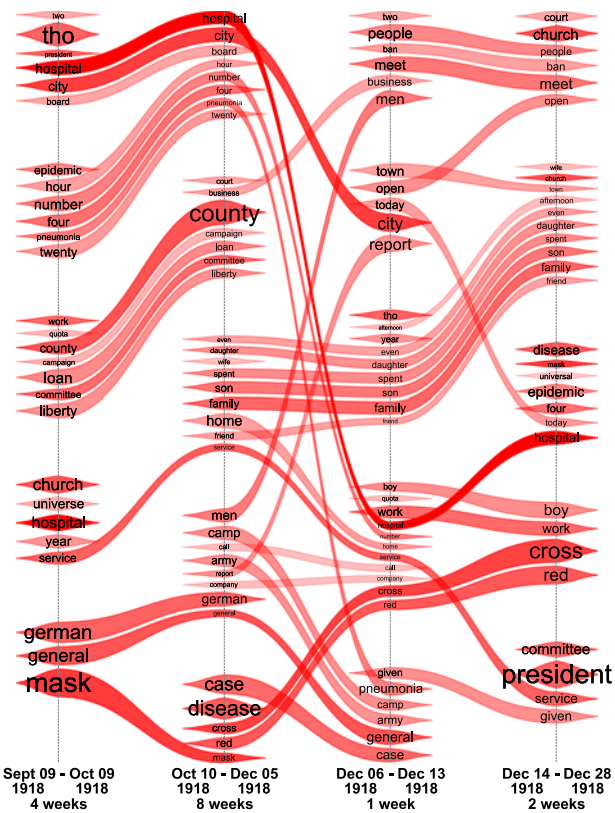



Fig. 10. ThemeDelta visualization for newspaper paragraphs during the period September to December in 1918. Color transparency for different trendlines signify the global frequency for that keyword.

the words "case," "disease," "mask," "cross," and "red" because the returning soldiers were exposed to the disease and some of them were sick. As a result, families were advised to take protective measures.

World War I ended on November 11, 1918, which explains the disappearance of the word "German," but the country continued suffering from the disease. The word "mask" reappeared back along with "epidemic," "hospital," and "disease" in the December 14 until December 28 segment, which aligns with the second influenza wave. Again, during this time people were advised to wear masks to slow down the spread of the disease. The Red Cross was frequently mentioned in the last three segments, which is indicative of the second, deadlier wave of the pandemic that began in October. In both the December 6 to December 13 and December 14 to December 28 segments, the terms "people," "ban," and "meet" appeared because people were banned from meeting each other as a precaution measure to limit the spread of the disease. The term "president" appeared in the last segment along with "service" appeared initially in the first segment and then returned with significant strength in the last segment, illustrating the seriousness accorded to the national scale of the pandemic.

## 9 Qualitative User Study

To validate the utility of the ThemeDelta system, including both its temporal segmentation algorithm as well as its visual representation, we conducted a qualitative user study involving expert participants. The purpose was to study the suitability of the approach for in-depth expert analysis of dynamic text corpora. Because of our existing collaboration with historians (the sixth author of this work is a historian), we opted to use the historical U.S. newspaper dataset and engage experts from the history department at one of our home universities.

### 9.1 Data

We used historical data from five U.S. newspapers for our qualitative evaluation from three different areas: New York, Washington, D.C., and Philadelphia. The data was collected from the Chronicling America website[5] and focused on the 1918 influenza epidemic, which killed as many as 50 million people worldwide and has long been recognized as one of the most deadly disease outbreaks in modern world history. Historians are interested in reconstructing the timeline of events, with a view to understanding previously concealed or neglected connections between public opinion, health alerts, and prevailing medical knowledge.

### 9.2 Method

We recruited three graduate students as participants: one from the history department and two from the English department at our university. The participants were all required to have prior knowledge of America around the Great War/First World War period. Two participants were Ph.D. students and one was a Masters student. We required no particular technical skill prior to participation. While the number of study participants may appear to be low, we want to emphasize that these participants represent a highly expert population and that our study protocol is focused more on an expert review [46] rather than a comparative or performance-based user study.

The total study time was an hour. The procedure was as follows: Participants were first asked to fill out a background questionnaire. Then the study moderator explained the tool and its features, followed by the task the participants were asked to perform using the tool. After that, the participants were asked to solve several high-level tasks (reviewed below) using the tool. Finally, they were asked to complete a post-session questionnaire to collect feedback on the tool.

The tasks that we asked the participant to accomplish with the help of our system was answering some questions on the 1918 influenza pandemic. Participants were encouraged to refer to the visualization in their answers by mentioning segments names, giving examples, or taking screen captures from the visualization. Tasks were divided into change and connection questions, to allow us to determine whether the visualization and algorithmic choices we made were helpful or not. The change-focused questions were:

- How did the newspapers describe the spread of influenza?
- How does the description of the pandemic change over time?
- Are there different times when the influenza pandemic becomes less important? What are those time periods?

Questions that were focused on connections were:

- What are the categories that appear to be associated with influenza in different newspapers?
- Was there a specific feeling that surrounded the influenza reporting in the newspapers?

### 9.3 Results

All three participants were successful in accomplishing the task using ThemeDelta. We determined this by comparing their answers to the task questions with model answers provided by the history faculty collaborator (reviewed in Section 8). They correctly reported the sentiments that surrounded the influenza from the five newspapers. They also successfully described the change in reporting of the influenza spread. Finally, they all succeeded in discovering the connection between influenza and other categories (e.g., schools, war, and hospitals).

The subjective results of the study were overall positive and the participants all vouched for the helpfulness of the system and the need for such systems in their research. None of the participants had previous experience using any visual analytics systems. This implies that the participants found ThemeDelta to be understandable and easy to use.

All the three participants finished the tasks within the allocated time. They also uniformly reported that the same type of task, if done manually as part of their own research, would normally take several days if not weeks. This highlights an additional strength to our system: minimizing the time spent on manual analysis of large amounts of text, allowing the analyst to focus on collecting insight instead.

5. Available online at: http://chroniclingamerica.loc.gov/

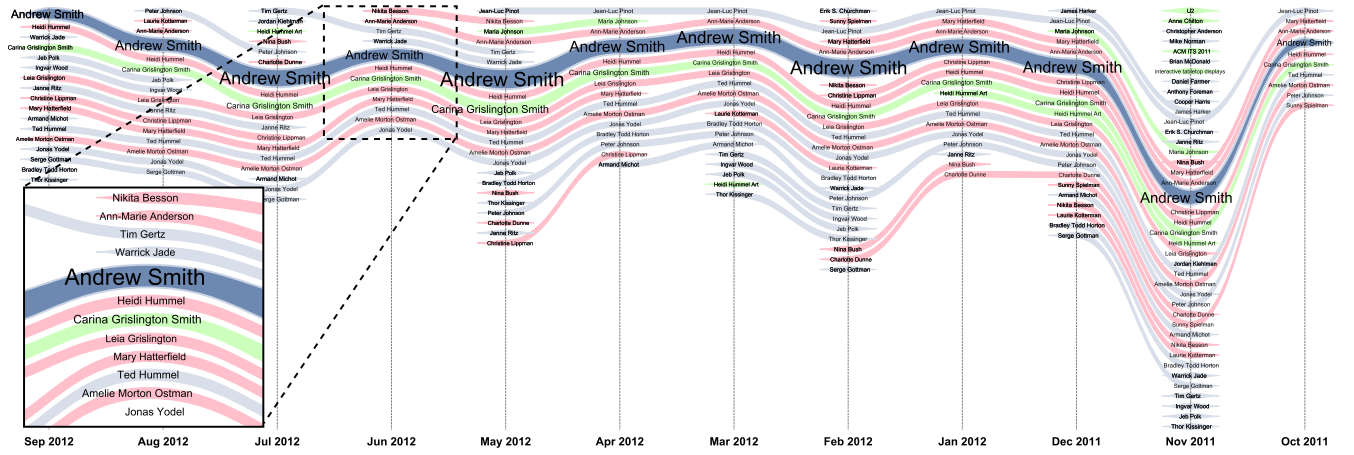



Fig. 11. FriendShare for Facebook app visualizing ego-centric communication patterns for the period October 2011 to September 2012 for a person named "Andrew Smith" (highlighted in dark blue). Trendlines representing male friends use a blue color, females use a pink color, and undisclosed gender (or entities) use green. The inset shows details for June 2012. Note the high number of connections in the month of November, Andrew's birthday month.

In the post-session questionnaire, participants were asked to give their feedback on specific ThemeDelta features. The features that were reported as very useful were labels, line thickness, duplicate trends, and discontinuations. Participant ratings for other features ranged from very useful to not useful at all, the latter typically because they did not use that particular feature. Some of the identified weaknesses of the tool included not being able to see full phrases or word combinations, managing keyword filtering, controlling the dynamic layout, and high complexity for large datasets.

## 10 Discussion

ThemeDelta excels at discovering trending topics and visualizing not just the discovered topics, but also their evolution over time. We have demonstrated the utility of our analytics component by applying it to several types of text corpora. Furthermore, we have also shown that the visualization part of ThemeDelta can be applied to other types of data beyond text; the FriendShare app is one example of how the technique can be used to show temporal edges in a dynamic graph, i.e., for edges that appear at specific points in time. However, while ThemeDelta has many strengths, it is also balanced by several weaknesses and areas of future improvement.

### 10.1 Visual Design

For the visualization component, limitations appear in the presence of many trends, long time periods, and high visual complexity. While existing techniques such as TextFlow [27] take a macroscopic approach to summarizing massive text corpora using high-level overviews, ThemeDelta uses a trend-level design that does not scale as well when the number of trends or time segments increases. While we have not derived a formal limit, even many of the examples presented in this paper skirt the boundary of the utility of the technique. In practice, large datasets (in either trends or time, or both) yield high visual complexity, particularly in the number of trendline crossings as well as incident trendlines. Such effects make perceiving the visualized data more difficult.

Several possible strategies can solve this problem, such as filtering, sampling, or aggregation. Fig. 11 shows FriendShare, a social media application for Facebook that we have developed to showcase one such strategy. FriendShare does not use the temporal segmentation algorithm proposed in this paper, but instead aggregates a Facebook user's interactions with friends in their social media network on a monthly basis. This data is collected and aggregated using the Facebook Graph API. Each friend becomes a trendline that appears and disappears for the periods when that friend interacts and does not interact, respectively, with the selected user. The example in Fig. 11 shows that "Andrew Smith" has several friends that regularly keep in touch with him, illustrated by the many unbroken trendlines in the visualization.

### 10.2 Topic Modeling

The underlying algorithm of ThemeDelta has limitations stemming from non-adaptive window sizing and fixed number of discovered topics. The minimum and maximum window sizes have to be pre-specified before the segmentation algorithm is run. This introduces the problem of force-adding a segmentation point when the maximum window size is reached, to overcome this we increase the maximum window size which can result in a slower running algorithm. The fixed number of topics discovered from each segment can introduce redundant topics. Overcoming this weakness effectively and systematically is a direction of future work.

### 10.3 Evaluation

We opted to *not* run a controlled quantitative experiment using ThemeDelta, Instead, we choose to run a qualitative expert review [46]. One reason for this choice is that we found no suitable technique to use as a baseline comparison for such an experiment. While techniques such as TextFlow [27] and TIARA [22] do provide insight on trends evolving over time, they cluster keywords together and focus on providing overview instead of detail at the level of individual keywords. In this sense, parallel tag clouds [23] are perhaps the closest technique to ThemeDelta in that it visualizes

individual keywords, yet PTCs do not show the clustering of trends into topics over time. This makes direct comparison difficult. While our qualitative review did not compare ThemeDelta to other techniques, it did give rise to much more qualitative and generally useful results.

## 11 Conclusion and Future Work

We have presented ThemeDelta, a visual analytics system for discovering and representing the evolution of trend keywords into ever-changing topic aggregations over time. While ThemeDelta builds on earlier text visualization techniques and text analytics algorithms, its main advantage is an emphasis on how keywords scatter and gather across a dataset. To showcase the applicability of ThemeDelta, we applied it to the political speeches of the two main candidates for the U.S. 2012 presidential campaign, social messaging practices in the i-Neighbors system, and historical U.S. newspaper data on the deadly Spanish Flu pandemic during the last part of 1918. We evaluated our system by conducting a qualitative expert user study using five different newspaper corpora. The study confirmed that ThemeDelta helped our expert participants in answering typical research questions in their fields.

A major future direction for this work is to extend ThemeDelta to be a complete visual analytics system wherein it supports knowledge generation as well. As mentioned in [47], and echoed in our own earlier work [48] this requires thoughtful integration of analyst's discovery processes through additional interfaces and metaphors. Our future work will also study aggregation methods—time-based and keyword-based alike—for ThemeDelta that would increase the scalability of the system. Another focus will be to enable the automatic detection and visualization of the diffusion of ideas in scientific communities. From the algorithmic perspective, we will be looking to extend the segmentation algorithm ability to work with nonparametric bayesian models as Hierarchal Dirichlet Process (HDP) proposed by Teh et al. [49]. Unlike LDA, number of topics in the HDP model is automatically inferred. Currently, our algorithm only supports parametric Latent Dirichlet allocation given that they are the most commonly used in practice and research. We also intend to continue to study applying additional advanced visualization and algorithmic techniques within the context of social websites such as Facebook.

## Acknowledgments

Special thanks to the qualitative user study participants.

## References

[1] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "ThemeRiver: Visualizing thematic changes in large document collections," *IEEE Trans. Vis. Comput. Graph.*, vol. 8, no. 1, pp. 9–20, Jan. 2002.

[2] L. Byron and M. Wattenberg, "Stacked graphs—geometry & aesthetics," *IEEE Trans. Vis. Comput. Graph.*, vol. 14, no. 6, pp. 1245–1252, Nov. 2008.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[4] i-neighbor website. [Online]. Available: http://i-neighbors.org/, 2007.

[5] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proc. IEEE Symp. Vis. Lang.*, 1996, pp. 336–343.

[6] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Reading, MA, USA: Addison-Wesley, 1999.

[7] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant, "Discovering interesting usage patterns in text collections: integrating text mining with visualization," in *Proc. ACM Conf. Inform. Knowl. Manage.*, 2007, pp. 213–222.

[8] M. A. Hearst and D. K. Rosner, "Tag clouds: Data analysis tool or social signaller?" in *Proc. Hawaii Int. Conf. Syst. Sci.*, 2008, pp. 160–160.

[9] S. Lohmann, J. Ziegler, and L. Tetzlaff, "Comparison of tag cloud layouts: Task-related performance and visual exploration," in *Proc.eedings of INTERACT*, ser. Lecture Notes in Computer Science, New York, NY, USA: Springer, vol. 5726, 2009, pp. 392–404.

[10] F. B. Viégas, M. Wattenberg, and J. Feinberg, "Participatory visualization with Wordle," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 6, pp. 1137–1144, Nov. 2009.

[11] K. Koh, B. Lee, B. H. Kim, and J. Seo, "ManiWordle: Providing flexible control over Wordle," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1190–1197, Nov./Dec. 2010.

[12] K. T. Kim, S. Ko, N. Elmqvist, and D. S. Ebert, "WordBridge: Using composite tag clouds in node-link diagrams for visualizing content and relations in text corpora," in *Proc. Hawaiian Int. Conf. Syst. Sci.*, 2011, pp. 1–8.

[13] F. van Ham, M. Wattenberg, and F. B. Viégas, "Mapping text with phrase nets," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 6, pp. 1169–1176, Nov. 2009.

[14] P. C. Wong, P. Mackey, K. Perrine, J. Eagan, H. Foote, and J. Thomas, "Dynamic visualization of graphs with extended labels," in *Proc. IEEE Symp. Inform. Vis.*, 2005, pp. 73–80.

[15] J. Clark. (2008, Oct.). Clustered word clouds. [Online]. Available: http://neoformix.com/2008/ClusteredWordClouds.html

[16] Y. Hassan-Montero and V. Herrero-Solana, "Improving tag-clouds as visual information retrieval interfaces," in *Proc. Int. Conf. Multidisciplinary Inform. Sci. Technol.*, 2006, pp. 25–28.

[17] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, "Visualizing the non-visual: Spatial analysis and interaction with information from text documents," in *Proc. IEEE Symp. Inform. Vis.*, 1995, pp. 51–58.

[18] B. Kwon, W. Javed, S. Ghani, N. Elmqvist, J. S. Yi, and D. Ebert, "Evaluating the role of time in investigative analysis of document collections," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 11, pp. 1992–2004, Sep. 2012.

[19] E. R. Tufte, *The Visual Display of Quantitative Information*. Cheshire, CT, USA: Graphics Press, 1983.

[20] M. Ghoniem, D. Luo, J. Yang, and W. Ribarsky, "NewsLab: Exploratory broadcast news video analysis," in *Proc. IEEE Symp. Vis. Anal. Sci. Technol.*, 2007, pp. 123–130.

[21] M. Wattenberg, "Visual exploration of multivariate graphs," in *Proc. ACM 2006 Conf. Human Factors Comput. Syst.*, 2006, pp. 811–819.

[22] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang, "TIARA: A visual exploratory text analytic system," in *Proc. ACM Conf. Knowl. Discov. Data Min.*, 2010, pp. 153–162.

[23] C. Collins, F. B. Viégas, and M. Wattenberg, "Parallel tag clouds to explore faceted text corpora," in *Proc. IEEE Symp. Vis. Anal. Sci. Technol.*, 2009, pp. 91–98.

[24] N. W. Kim, S. K. Card, and J. Heer, "Tracing genealogical data with timenets," in *Proc. ACM Conf. Adv. Vis. Interfaces*, 2010, pp. 241–248.

[25] C. Turkay, J. Parulek, N. Reuter, and H. Hauser, "Interactive visual analysis of temporal cluster structures," *Comput. Graph. Forum*, vol. 30, no. 3, pp. 711–720, 2011.

[26] W. Dou, X. Wang, R. Chang, and W. Ribarsky, "ParallelTopics: a probabilistic approach to exploring document collections," in *Proc. IEEE Conf. Vis. Anal. Sci. Technol.*, 2011, pp. 231–240.

[27] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong, "TextFlow: Towards better understanding of evolving topics in text," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2412–2421, Nov. 2011.

[28] A. Hotho, A. Nürnberger, and G. Paass, "A brief survey of text mining," *LDV Forum*, vol. 20, no. 1, pp. 19–62, 2005.

[29] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. Conf. Uncertainty Artif. Intell.*, 1999, pp. 289–296.

[30] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 113–120.

[31] X. Wang and A. McCallum, "Topics over time: a non-Markov continuous-time model of topical trends," in *Proc. ACM Conf. Knowl. Discov. Data Mining*, 2006, pp. 424–433.
[32] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda, "Online multiscale dynamic topic models," in *Proc. ACM Conf. Knowl. Discov. Data Min.*, 2010, pp. 663–672.
[33] C. Wang, D. Bleid, and D. Heckerman, "Continuous time dynamic topic models," in *Proc. Conf. Uncertainty Artif. Intell.*, 2008, pp. 579–586.
[34] G. F. Lawler, *Introduction to stochastic processes*. Boston, MA, USA: Chapman & Hall/CRC, 1995.
[35] X. Wang, S. Liu, Y. Song, and B. Guo, "Mining evolutionary multi-branch trees from text streams," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2013, pp. 722–730.
[36] X. Wang, K. Zhang, X. Jin, and D. Shen, "Mining common topics from multiple asynchronous text streams," in *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, ser. WSDM '09. New York, NY, USA: ACM, 2009, pp. 192–201.
[37] X. Wang, C. Zhai, X. Hu, and R. Sproat, "Mining correlated bursty topic patterns from coordinated text streams," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '07. New York, NY, USA: ACM, 2007, pp. 784–793.
[38] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09. New York, NY, USA: ACM, 2009, pp. 497–506.
[39] Z. Gao, Y. Song, S. Liu, H. Wang, H. Wei, Y. Chen, and W. Cui, "Tracking and connecting topics via incremental hierarchical Dirichlet processes," in *Proc. Data Min. IEEE 11th Int. Conf.*, 2011, pp. 1056–1061.
[40] S. Gad, N. Ramakrishnan, K. N. Hampton, and A. Kavanaugh, "Bridging the divide in democratic engagement: Studying conversation patterns in advantaged and disadvantaged communities," in *Proc. IEEE Conf. Soc. Inform.*, 2012, pp. 165–176.
[41] N. Elmqvist and P. Tsigas, "Causality visualization using animated growing polygons," in *Proc. IEEE Symp. Inform. Vis.*, 2003, pp. 189–196.
[42] I. Davidson, L. T. Watson, M. S. Hossain, and N. Ramakrishnan, "How to alternatize a clustering algorithm," in *Proc. Data Min. Knowl. Discov.*, vol. 27, no. 2, pp. 193–224, 2013.
[43] G. DiBattista, P. Eades, R. Tamassia, and I. G. Tollis, *Graph Drawing: Algorithms for the Visualization of Graphs*. Englewood Cliffs, NJ, USA: Prentice Hall, 1998.
[44] Y. Tanahashi, and K.-L. Ma, "Design considerations for optimizing storyline visualizations," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2679–2688, Oct. 2012.
[45] S. Liu, Y. Wu, E. Wei, M. Liu, and Y. Liu, "Storyflow: Tracking the evolution of stories," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2436–2445, Dec. 2013.
[46] M. Tory and T. Möller, "Evaluating visualizations: Do expert reviews work?" *IEEE Comput. Graph. Appl.*, vol. 25, pp. 8–11, Sep. 2005.
[47] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim, "Knowledge generation model for visual analytics." in *Proc. IEEE Trans. Vis. Comput. Graph*, vol. 20, no. 12, 2014, pp. 1604–1613.
[48] A. Endert, M. Hossain, N. Ramakrishnan, C. North, P. Fiaux, and C. Andrews, "The human is the loop: New directions for visual analytics," in *Proc. J. Intell. Inform. Syst.*, 2014, pp. 1–25.
[49] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," in *Proc. J. Am. Stat. Assoc.*, vol. 101, no. 476, pp. 1566–1581.

**Samah Gad** received the Bachelor and Master degrees in computer engineering from the Arab Academy for Science and Technology in Alexandria, Egypt, and the PhD degree from Virginia Tech in Blacksburg, VA, in 2014. She is currently the chief technology officer and Co-Founder of KustomNote a SaaS company. Her research interest is data mining, visual analytics, and machine learning. She is a member of the IEEE.

**Waqas Javed** received the Bachelor of Science in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2007, and the PhD degree from Purdue University in West Lafayette, IN, in 2013. Currently he is working as a HCI researcher at General Electric in San Ramon, CA. His research interests include information visualization, visual analytics, and human-computer interaction. He is a member of the IEEE.

**Sohaib Ghani** received the Bachelor of Science in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2007, and the PhD. degree from Purdue University in West Lafayette, IN, in 2013. Currently he is working as an assistant professor at KACST GIS Technology Innovation Center, Umm Al-Qura University, Makkah, Saudi Arabia. His research interests include information visualization, and visual analytics. He is a member of the IEEE.

**Niklas Elmqvist** received the PhD degree from Chalmers University of Technology in Göteborg, Sweden, in 2006 . Currently, he is an associate professor in the College of Information Studies at University of Maryland, College Park, MD. He was previously an assistant professor in the School of Electrical & Computer Engineering at Purdue University in West Lafayette, IN. Before that he was a postdoctoral researcher in the Aviz group at INRIA Saclay in Paris, France. He is a senior member of the IEEE.

**Tom Ewing** received the BA degree from Williams College, Williamstown, MA, and the PhD degree in history from the University of Michigan, Ann Arbor, MI. He currently the associate dean for Graduate Studies, Research, and Diversity in the College of Liberal Arts and Human Sciences and professor of History at Virginia Tech, Blacksburg, VA. His articles on digital history have appeared in Perspectives on History, Computer (IEEE), Social Education, and The Journal of Women's History.

**Keith N. Hampton** received the BA degree in sociology from the University of Calgary, Calgary, AB, Canada, in 1996, and the MA and PhD degrees from the University of Toronto, Toronto, ON, Canada, in 2001 and 1998, respectively. He is currently an associate professor in the Department of Communication at Rutgers University, New Brunswick, NJ. His research interests focus on the relationship between new information and communication technologies, social networks, democratic engagement, and the urban environment.

**Naren Ramakrishnan** received the PhD degree in computer sciences from Purdue University, West Lafayette, IN, in 1997. He is currently the Thomas L. Phillips Professor of Engineering in the Department of Computer Science at Virginia Tech, Blacksburg, VA. His research has been supported by NSF, DHS, NIH, NEH, DARPA, IARPA, ONR, General Motors, HP Labs, NEC Labs, and Advance Auto Parts. He has served as both program chair and general chair of the IEEE International Conference on Data Mining (ICDM). He is a member of the IEEE.