Visual Analysis of Topic Competition on Social Media

Panpan Xu, Yingcai Wu, *Member, IEEE*, Enxun Wei, Tai-Quan Peng,
Shixia Liu, *Senior Member, IEEE*, Jonathan J. H. Zhu, and Huamin Qu, *Member, IEEE*

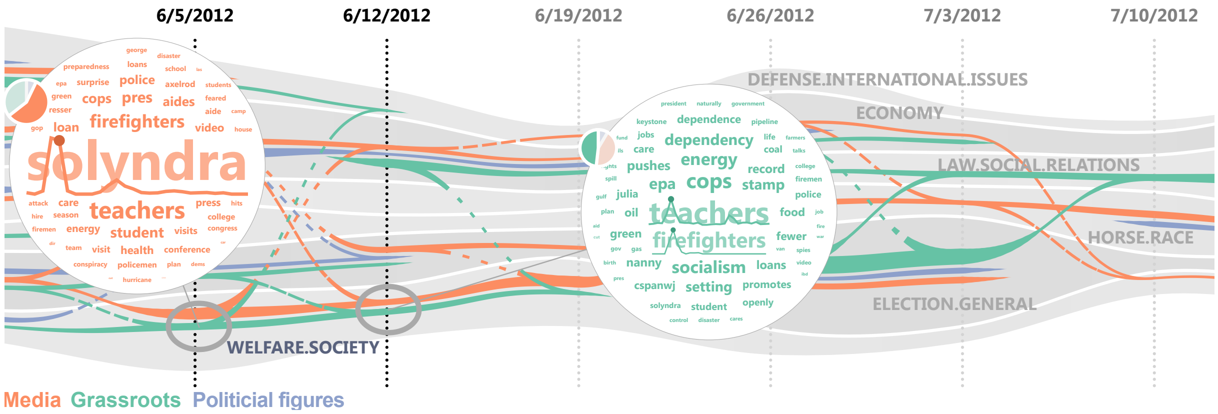

Fig. 1. The *grassroots* and the *media* played the major roles in recruiting public attention to the topic *welfare / society* on Twitter. The trending keywords related to the topic were not the same at different times. Around June 5th, there was a trending discussion on *solyndra*, while around June 12th, the keywords "teachers" and "firefighters" gained more significance. The sparklines underlying the keywords have their peaks on the selected time interval, indicating potential correlation between those topics to the recruitment effects observed. The tweets containing the keywords can also be examined to find the actual events that trigger the discussions.

Abstract—How do various topics compete for public attention when they are spreading on social media? What roles do opinion leaders play in the rise and fall of competitiveness of various topics? In this study, we propose an expanded topic competition model to characterize the competition for public attention on multiple topics promoted by various opinion leaders on social media. To allow an intuitive understanding of the estimated measures, we present a timeline visualization through a metaphoric interpretation of the results. The visual design features both topical and social aspects of the information diffusion process by compositing ThemeRiver with storyline style visualization. ThemeRiver shows the increase and decrease of competitiveness of each topic. Opinion leaders are drawn as threads that converge or diverge with regard to their roles in influencing the public agenda change over time. To validate the effectiveness of the visual analysis techniques, we report the insights gained on two collections of Tweets: the 2012 United States presidential election and the Occupy Wall Street movement.

Index Terms—Social media visuaization, topic competition, information diffusion, information propagation, agenda-setting

1 Introduction

As an influential theory in mass communication research, agenda-setting asserts that the emphasis of certain topics (issues) in news media determines their saliency as perceived by the general public [28]. In other words, by telling people what to think about, news media influences the salience of topics on the public agenda. Nowadays, news media are not the only agenda setters in society. The advent of social media, such as Twitter and Facebook, has empowered ordinary users to influence media emphasis and the perceived saliency of certain topics among the general public. Social media users can generate voluminous information and disseminate them to a huge number of people, thus having the opportunity to influence the saliency of a topic on the public agenda. Almost 90% of public relations practitioners believe that social media has become a new form of media to set the public agenda [47].

The Egyptian Revolution of 2011 is an example of how Twitter users employed social media to establish the agenda of the news media and the public. When the revolution started, Americans paid little attention to the event. To seek support, revolution leaders started promoting topics, such as "#jan25", on Twitter, which spread quickly and attracted the attention of the mass media. United States (US) President Barack Obama subsequently showed his support to the revolution through a speech, which significantly increased media attention on the event. Ultimately, the Egyptian Revolution became popular news worldwide and salient on the public agenda.

The rapid development of social media does not only provide new opportunities, but also poses challenges in agenda-setting. The abundance of information available on social media exceeds the limited carrying capacity of the public agenda [45, 50]. As such,topics have to compete for the scarce public attention [50]. The saliency of topics is directly influenced by the simultaneously available competing topics [4, 48]. The more competing topics on social media, the less salient an individual topic is likely to be.

Moreover, prior research reveals that most people on social media acquire information from a group of elite users (i.e., opinion leaders), based on the two-step flow of communication theory [17, 49]. Opinion

- *Panpan Xu and Huamin Qu are with Hong Kong University of Science and Technology. E-mail: {pxu, huamin}@cse.ust.hk.*
- *Yingcai Wu and Shixia Liu are with Microsoft Research Asia. Y. Wu is the correspondence author. E-mail: {yingcai.wu, shixia.liu}@microsoft.com.*
- *Enxun Wei is with Shanghai Jiao Tong University. E-mail: weienxun@gmail.com.*
- *Tai-Quan Peng is with Nanyang Technological University. E-mail: winsonpeng@gmail.com.*
- *Jonathan J.H. Zhu is with City University of Hong Kong. E-mail: j.zhu@cityu.edu.hk.*

Manuscript received 31 March 2013; accepted 1 August 2013; posted online 13 October 2013; mailed on 4 October 2013.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

leaders act as intermediate layers through which information is filtered and interpreted based on their own perceptions. They can potentially alter the saliency of a topic as perceived by other users; thus, opinion leaders play a gate-keeping role in the agenda-setting process. The competitive relations among topics and the involvement of opinion leaders complicate the dynamics of agenda-setting, which is rarely addressed in current research.

In the context of agenda-setting on social media, the dynamics of topic competition occur at a larger scale with multiple topics and among different groups of opinion leaders. In particular, multiple types of time-varying relations accompany the agenda-setting process, including the competition among topics and the agenda-setting effect of different groups of opinion leaders on the topics. Dynamic multiple-type relationships compound the difficulties in creating a concise and readable visual representation. Apart from providing an overview of the evolving relations, seeking explanations for the ebb and flow of the competitiveness of topics, and formulating interpretations for the commonalities and differences among the agenda-setting effects of different opinion leaders are challenging. These tasks require the system to provide a mechanism for investigative analysis, for which the contents of tweets must be made accessible. Currently available systems [41, 8] are limited to a single post or topic. To our knowledge, none of the existing studies has helped analyze agenda-setting and topic competition effects on social media, particularly the roles played by opinion leaders in the dynamics of agenda-setting.

In this study, we propose a visual analysis framework to study agenda-setting and topic competition effects on social media. Unlike previous literature on social media analytics, our approach analyzes the following patterns on social media: (1) the dynamic competition among various topics to gain public attention and (2) the roles that opinion leaders play in the process. We designed a timeline visualization to provide an overview of the competitiveness and saliency of topics over time, the commonalities and differences of the roles played by opinion leaders in the agenda-setting process, and the transition of topical focus of opinion leader groups. Based on the (temporal) correlation patterns revealed in the overview, several hypotheses are formed. Investigative analysis and result validation are supported by word clouds. To illustrate the usefulness of our approach, we applied visual and analytical methods on two collections of tweets: the 2012 U.S. presidential election and the Occupy Wall Street movement. We also interviewed sociologists and provided a summary of their feedback.

The contributions of the study presented in this paper include:

- An expanded topic competition model of agenda-setting to characterize the dynamics of topic competition and the roles played by opinion leaders.
- A set of techniques for visualizing the temporal and heterogeneous relationships identified by the expanded model.
- Two case studies to explore the complex dynamics of agenda-setting and topic competition on social media.

2 Related Works

This section briefly discusses related literature on information diffusion, social media visualization, and temporal data visualization.

2.1 Information Diffusion

Various models, such as the linear threshold model and the independent cascade model [13], have been used to characterize the diffusion of information on social networks. However, the models view the diffusion of each topic to be independent. Multiple topics may influence each other when they are disseminated on social networks. Several studies have also shown that increasing the number of competing topics results in decreasing popularity of the topics [4, 48].

Researchers have only recently studied the diffusion of multiple topics. Myers and Leskovec [31] developed a probabilistic model to reflect the interactions of different topics when they are diffused in the network. Weng et al. [45] proposed a parsimonious agent-based model to examine the relationship between competition and the popularity, diversity, and lifetime of a topic. Our work also studies the problem of characterizing the diffusion of multiple topics. Compared with existing research, this study models the diffusion process with no assumption about the existence of an underlying network. It also analyzes the competition for the aggregated public attention (i.e., how the opinion leaders set the public agenda), instead of modeling how individual users are affected by the contagions. Moreover, we derive metrics that characterizes the time varying competiveness of temporal salient topics (e.g., economics), different from the related research that models the diffusion of more volatile contagions (i.e., memes). We also identify and evaluate the roles of different groups of opinion leaders in the process of setting the public agenda.

2.2 Social Media Visualization

Researchers have developed various visualization methods to help users understand social media data by providing aggregated information [10, 11, 26].Diakopoulos et al. [10] presented a visual analytics tool with multiple linked views to aid journalists and media professionals in analyzing social media content. Marcus et al. [26] developed TwitInfo to automatically detect and display peaks of high tweet activity. Dörk et al. [11] introduced a web-based system that can provide a visual summary of large-scale Twitter data streams.

Previous research has also used clustering to reduce data complexity and to facilitate analysis [3, 14]. Gansner et al. [14] described a text stream visualization method that initially groups tweets by “countries” and then generates a dynamic map. ThemeCrowds[3] displays topic trends on Twitter over time using multi-scale tag clouds. Twitter users are clustered hierarchically and then visualized based on the topics they discuss. Recently, several visualizations [41, 1, 2] have been designed to show the spread of information on social media. Nan et al. [8] developed Whisper to visualize the spatio-temporal process of information diffusion on Twitter. Nevertheless, these systems mainly focus on visualizing the diffusion process of a typical event on social media. Visualizing the simultaneous dissemination of multiple events using these systems is extremely difficult, if not impossible. By contrast, our study aims to visualize the dynamics of multiple competing topics.

2.3 Temporal Data Visualization

There have been various approaches for visualizing and analyzing temporal data. Comprehensive surveys could be found in [21] [29]. Researchers have extended visualization techniques such as parallel coordinate plots for visualizing time variant multivariate data [18] and histograms for time varying data distributions [20]. Time is often represented by the horizontal axis [21, 29], or spirals to highlight the periodical patterns [44]. Aggregation by clustering similar time series has been used to visualize a large number of time series [40]. In this work, we analyze a large volume of tweets, which are semantically rich textual data linked to the temporal dimension. We attempt to unearth the dynamics of agenda-setting and topic competition process from the data by applying time series modeling techniques and visualize the time varying metrics and relations based on the estimated model. We discuss the most related temporal data visualization techniques below.

ThemeRiver [16] is the first system to automatically create a smooth stacked graph layout that can handle many time series. The approach has been adopted by NameVoyager [43] to visualize baby name popularity. Byron and Wattenberg [7] introduced methods to create aesthetically appealing stacked graph layouts using optimization. Recently, researchers have also extended Streamgraph to support visual analysis of large-scale text corpora [9, 11, 24, 38]. Rose et al. [35] introduced “story flow” visualization for tracking the evolution of themes in text streams. Cui et al. [9] introduced TextFlow to show the relationships among topics in text corpora. RankExplorer [37] adds color bars and glyphs to stacked graphs to display changes in item values and their ranking. EventRiver [25] extracts and visualizes the events within text collections with temporal reference. Outflow [46] applied sankey diagram to visualize temporal event sequence.

Storyline visualization, as inspired by Munroe’s movie narrative charts [30], can intuitively convey relationships among entities over time. PlotWeaver [33] is a Web-based system that allows users to interactively create storylines from scratch. Ogawa and Ma [32] presented a set of design criteria and derived a greedy algorithm to create a layout that satisfied the criteria. Yuzuru and Ma [39] refined the design criteria and used a genetic algorithm to generate an aesthetically-appealing and legible storyline layout. A different layout algorithm has also been proposed by Liu et al. [23]. The storyline style visualization has also been applied in different applications such as genealogical data analysis [19] and community detection [34].

Our timeline method also conveys converging and diverging behaviors among entities, which is similar with storyline visualization. However, opinion leaders can exert influence on multiple topics at a time interval and their topical focus can shift among multiple topics as time progresses. Therefore, the dynamic relation that can be expressed using storylines needs to be extended in our case. We subsequently employ a composite visual design that integrates the dynamics of the influences of opinion leaders with the ebb and flow of competitiveness and saliency of topics. The composite style is similar with TextFlow [9] which draws keywords as threads over sankey graphs to depict their co-occurrence in time-varying topic clusters. In our case, the branching and merging of threads on ThemeRiver encode the semantic meaning of topical transition, which is different from that on TextFlow.

3 SYSTEM OVERVIEW

The visual analysis framework is illustrated in Figure 2. The framework has three major components: *data storage and preprocessing*, *data analysis*, and *interactive visualization*. The data preprocessing component employs Apache Lucene, a high-performance text search engine [1], to enable text indexing and searching. This search engine library allows the data analysis component to extract time-series data efficiently to model topic competition. It also supports dynamic query as user interacts with the visualization.

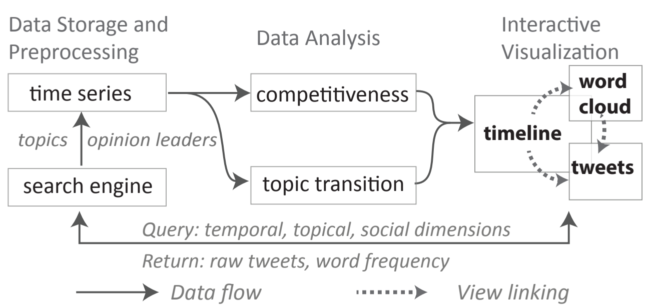

Fig. 2. System overview. The system consists of three major parts: data storage and preprocessing, data analysis, and interactive visualization.

The data analysis component characterizes the dynamic competition of several important topics, such as economy and welfare, based on an expanded competition model. More importantly, this component can quantitatively estimate the influence of an opinion leader group, such as "media" and "politicians", on each topic for each time interval using the aforementioned model. The topic transition trend of each opinion leader group (that is, the degree of their attention switching from one topic to another) is also evaluated by the component for each time interval.

The visualization component accepts input from the analysis model, including: (1) the competitiveness of each topic over time, (2) the influence exerted by an opinion leader group on the competitiveness of each topic over time, and (3) the trend of topic transition over time. The data can be regarded as the co-evolutionary relations between topics and opinion leader groups. The component employs a timeline design which overlays threads, which represent opinion leaders, on a stacked graph, which represents topics, to intuitively reveal co-evolutionary relations. Analysts can select any time interval from the timeline visualization to perform investigative analysis through details-on-demand. In particular, a radial graph is used to reveal pairwise competition relation among the topics for the selected time interval. A more in-depth investigative analysis of the tweet content is supported by displaying a word cloud within the radial graph.

[1]http://lucene.apache.org/core/

4 MODELING TOPIC COMPETITION ON SOCIAL MEDIA

In this section, we introduce background knowledge on agenda-setting theory and briefly describe the original topic competition model applied in the study of mass media [50]. After that, we propose the expanded model with multiple influence sources, such that the effect of different opinion leader groups could be differentiated for the study of the topic competition on social media. Based on the expanded quantitative model, we derive some intuitive measures on the competitiveness of a topic and on the parts that are contributed by each opinion leader group. The resulted measures enable our visualization system to create a comprehensive view of the dynamic competition process.

4.1 Agenda-setting and Topic Competition

4.1.1 Media vs. Public Agenda

Agenda-setting research starts with the definition of two agendas, called media agenda and public agenda, respectively. The former refers to a set of topics that are prominently reported by the mass media whereas the latter refers to a set of topics that are considered important by majority of the public in a society. Agenda-setting research focuses on the causal relationship between the two sets of topic salience. In a seminal study, McCombs and Shaw [28] found a strong correlation between media agenda and public agenda. They call this correlation as an agenda-setting effect, suggesting that the media set the agenda for the public by emphasizing certain topics in news coverage while downplaying others.

4.1.2 Topic Competition

In agenda-setting, multiple topics have to compete for media coverage and public attention, as the addition of any new topic onto the public agenda comes at the cost of other topic(s) [50]. The competition among the topics is necessarily caused by structural constraints on a wide range of stakeholders, including the limited capacity of the public to process information, the limited space or time of the media to cover news events, and the limited attention or resources of politicians, interest groups, and the entire social system at large to deal with competing topics. Zhu [50] developed a difference equations system (Equation 1) to model two competition mechanisms: 1) recruitment effect for attracting followers of topic j to i and 2) defection effect for distracting followers of i to j.

$$\Delta p_i^t = m_i^{t-1} \sum_{j=1, j\neq i}^{k} \beta_{ij} p_j^{t-1} + p_i^{t-1} \sum_{j=1, j\neq i}^{k} \beta_{ji} m_j^{t-1}, \text{ for } \forall i \in \{1, \ldots, k\} \quad (1)$$

In Equation 1, the independent variables include m_i^{t-1}, the media coverage on topic i at $t-1$, and p_i^{t-1}, the perceived salience of topic i by the public at $t-1$. The parameters to be estimated are β_{ij}. The dependent variable Δp_i^t describes the change in the perceived salience of topic i by the public, as measured by the difference $(p_i^t - p_i^{t-1})$ between the proportion of the public considering i to be important at two adjacent time points t and $t-1$. The change is assumed to be caused by the two competing forces as described by the terms on the right side of Equation 1: the recruitment effect by media coverage of i (i.e., m_i^{t-1}) on the followers of j (i.e., p_j^{t-1}) and the defection effect by media coverage of j (i.e., m_j^{t-1}) on the followers of i (i.e., p_i^{t-1}). There are k topics and therefore k parallel equations in total. Assuming both media coverage and public perception of the topics are measured along discrete time points, which is usually the case, the parameters β_{ij} and β_{ji} can be directly estimated by a standard regression system.

An empirical test of the model with three public topics in the U.S. in 1990-91 shows three possible outcomes of topic competition: one-way attraction, mutual competition, and independent coexistence [50]. Subsequent studies have shown a variety of long-term consequences of topic competition. For example, McCombs and Zhu [27] found that the competition among topics becomes increasingly tough over time, which leads to a faster rate of topic turnover on the public agenda.

4.2 Expanded Model: Multiple Influence Sources

The model as described by Equation 1 assumes homogeneity among the mass media agendas, thus treating them as a whole as they influence the competition among the topics and consequently the shaping of public agenda. However, in the social media setting, there could be considerable diversity in the agendas of different opinion leaders and in their impact on the public agenda. Therefore, a fine-grained model which differentiates various types of opinion leaders is desirable for the study of agenda-setting and topic competition effect on social media. The original model is thus expanded, where the term m_i^{t-1}, representing the overall media coverage on topic i at time $t-1$, is replaced with terms representing the coverage on the topics of different opinion leader groups (i.e., $m_{i,g}^{t-1}$ in Equation 2). With the new terms, we could model the recruitment effect by opinion leader group g's coverage on topic i on the followers of j and the defection effect by opinion leader group g's coverage on topic j on the followers of i. An autoregressive term is also added to account for the carry over effect from the last time point.

$$p_i^t = \alpha_i p_i^{t-1} + \sum_{g=1}^{n} m_{i,g}^{t-1} \sum_{j=1,j\neq i}^{k} \beta_{i,j,g} p_j^{t-1} - p_i^{t-1} \sum_{j=1,j\neq i}^{k} \Big(\sum_{g=1}^{n} \beta_{j,i,g} m_{j,g}^{t-1}\Big) \quad (2)$$

Figure 3 is a conceptual diagram which illustrates the model by highlighting the engagement of two opinion leader groups (marked with different colors) in topic i (i.e., m_{i,g_0} and m_{i,g_1}) and their effect in drawing public attention away from topic j. The edges between topic i and j correspond to the recruitment effect by opinion leader groups' engagement in topic i on the followers of topic j, which is reflected in the model as product terms $m_{i,g_0} p_j$ and $m_{i,g_1} p_j$. In the original model, mass media is treated as a single agenda setter.

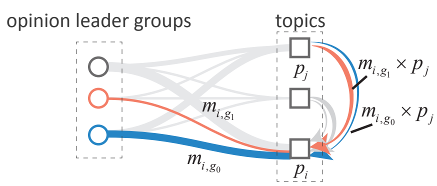

Fig. 3. Illustration of the model with two groups of opinion leaders' engagement in topic i drawing the attention of the public away from topic j.

4.3 Measuring Competitiveness

The model in Equation 1 and the expanded model in Equation 2 for social media can be estimated by treating the product terms of the independent variables (e.g. $m_i^{t-1} p_j^{t-1}$ and $m_{i,g}^{t-1} p_j^{t-1}$) as individual independent variables and solving a (consequently) linear regression system. However, it is still needed to derive reliable and intuitively interpretable quantitative measures for the competitiveness of the topics and the recruitment effect that could be attributed to each opinion leader group, such that they could be mapped to visual variables.

Measuring recruitment effects. Conceptually, the competitiveness of topic i refers to the total effect size of the recruitment effects (i.e., $m_{i,g}^{t-1} p_j^{t-1}$) by the opinion leader groups advocating i on the followers of other competing topics. How then will recruitment effects be quantitatively measured? The answer is not as straightforward as it appears. It is not because of the lack of relevant parameters from empirical tools (e.g., regression analysis). On the contrary, the difficulty rises from the existence of several measures of effects, each with its strengths and shortcomings, such as unstandardized regression coefficient (b), standardized regression coefficient (β), squared partial correlation (pr^2), and squared semipartial correlation (sr^2). Of the parameters, we choose to use sr^2 to measure topic competitiveness in the current study because sr^2 is the most stringent measure by describing the unique proportion of the variance in the dependent variable (p_i^t in Equation 2) that is solely explained by an independent variable (i.e., $m_{i,g}^{t-1} p_j^{t-1}$ in the context of Equation 2) after removing not only the unique variance explained by competing independent variable (e.g., $m_{j,g}^{t-1} p_i^{t-1}$) for the defection effect but also the joint variance explained by all independent variables due to the multicollinearity among them.

As such, sr^2 is a conservative measure that denotes the lower bound of the effect size of any independent variable (i.e., $m_{i,g}^{t-1} p_j^{t-1}$). However, the unique-effect nature makes sr^2 to possess two desirable properties: it is additive so that sr^2 from multiple equations of an equation system for a given independent variable can be summed to form an overall measure of effect size; and it is normalized so that they are directly comparable within an equation or over all equations of a system.

Measuring the effect of multiple influence sources and topics. Since sr^2 as measure for the effect of individual independent variables is additive and comparable, it serves our purpose particularly well, as illustrated with the following procedure to obtain quantitative measures for the competitiveness of each topic and the part that is contributed by each opinion leader group:

For each of the recruitment terms (i.e., $m_{i,g}^{t-1} p_j^{t-1}$) in Equation 2, we can obtain a corresponding piece of sr^2, denoted as $sr_{i,j,g}^2$. Summing the pieces with the same subscript g (i.e., $\sum_{j=1,j\neq i}^{k} sr_{i,j,g}^2$), we obtain the competitiveness contributed by opinion leader group g to topic i (i.e., the recruitment effect of opinion leader group g with respect to topic i); summing the pieces with the same subscript i (i.e., $\sum_{g=1}^{n} \sum_{j=1,j\neq i}^{k} sr_{i,j,g}^2$), we obtain the overall competitiveness for each of the k topics; summing the pieces with the same subscript i and j, we obtain the pairwise competitiveness among the topics.

The resulting measures of topic competitiveness provides the empirical basis for comparisons, analytically or visually, among the n groups of opinion leaders within each topic over time, or among the k topics across all groups of opinion leaders over time. The measures of *competitiveness* and *recruitment effect* all characterize the agenda-setting process. They describe the mechanism of how the topic agendas of the opinion leaders shape the public agenda by transferring the public attention among the topics, increasing the popularity of some while deemphasizing others. We will also use the terms such as "attract", "influence" to describe *recruitment effect* in the rest of the paper as they are more intuitive to interpret.

5 Visual Design

This section briefly discusses the design process and the concrete design goals , and then describes the interactive visual analysis system.

5.1 Design Process

We work closely with two domain experts on media study who are also co-authors of this paper. A wide range of decisions crucial for data analysis and modeling, and for the subsequent visual designs are formulated through frequent exchange of opinions and extensive experimentation on both sides. These include the choice of the regression model, the topics and opinion leader groups that are involved in the analysis. The discussions have gradually lead to a concrete project with clearer design goals. An iterative process is also adopted in the design of the system: an initial prototype with synthetic data, and a number of subsequent visual design mockups have been demonstrated to collaborators to gather feedback.

5.2 Design Goal

The visualization should display the competition for public attention among multiple topics promoted by multiple types of opinion leaders in agenda-setting on social media, based on quantitative measures obtained from the expanded model (Equation 2). The quantitative measures include the time varying competitiveness of each topic, the contribution made by each opinion leader group to the competitiveness of the topics, and the pairwise competition among the topics. The system should also enable in-depth analysis to gain insight into the potential causes of the observed competition effect. We have identified a set of research problems , which are listed as follows:

Q1 Which opinion leader groups contribute most to the competitiveness of a topic?

Q2 What are the commonalities and differences among opinion leaders with regard to their time-varying influences on the public agenda? How do their influences converge and diverge over time?

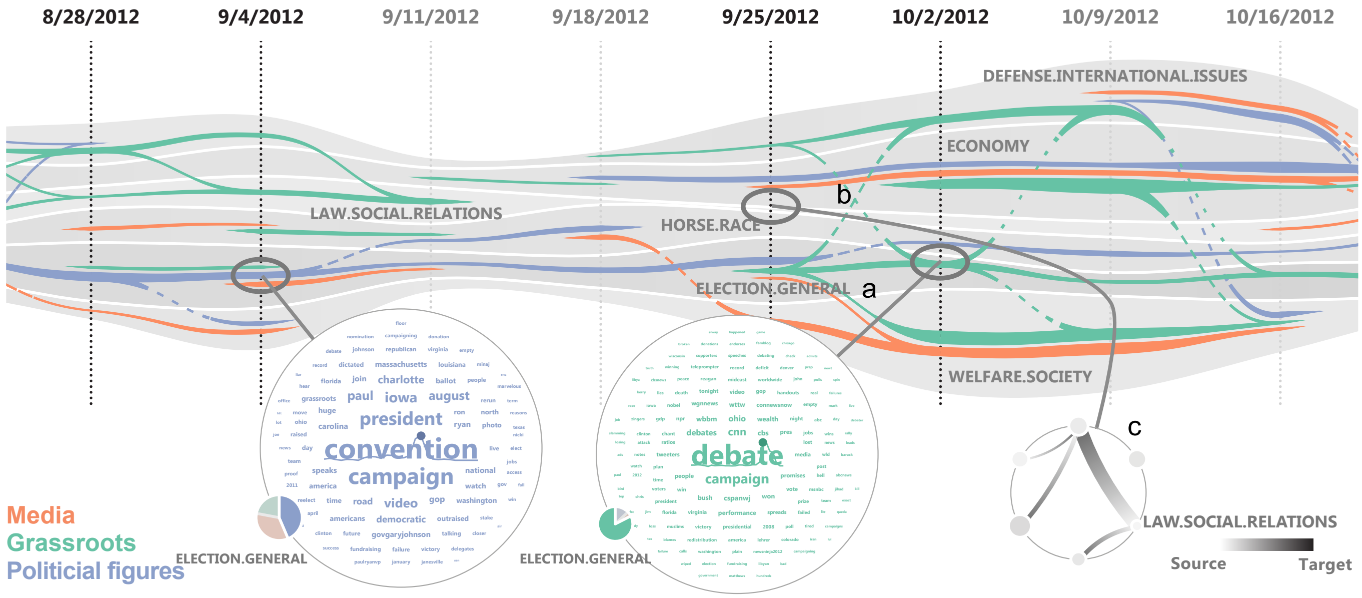

Fig. 4. In Election data, interesting patterns were observed around August 28th and September 4th, which could be related to *Republican and Democratic national conventions*. Other patterns could also be observed around the time of *Presidential debates*.

Q3 How do the influence and topical focus of opinion leaders evolve? How do they co-evolve with the competitiveness of the topics?

Q4 How do the topics compete with each other pairwisely?

Q5 Given the temporal trend in the rise and fall of the competitiveness of the topics, could plausible explanations be formed? For example, is it possible that the occurrence of an event would trigger an increase in the competitiveness of a topic?

Q6 For the evolutionary and co-evolutionary patterns discovered, could preliminary hypotheses be formed on the possible causes? The contents of tweets could be analyzed to generate explanatory hypotheses for the patterns, such as the simultaneous influence of multiple opinion leader groups on the same topic, the absence of influence of some opinion leaders, and the transition of interest of the opinion leaders.

The above questions have framed the scope and serve as inspirations for the visual designs. They also pose challenges on the design of effective visual presentations. For example, to enable efficient temporal correlation pattern detection (Q1, Q2, Q3), our collaborators have intended for the related data to be integrated in one view. These data include the time varying competitiveness and saliency of the topics (Q1, Q3), the competitiveness contributed by each group to the topics (Q1, Q2 and Q3), and the transition of topical focus of each opinion leader group (Q3). Although each piece of information can be displayed separately using existing visualization techniques (such as line chart and ThemeRiver for multiple time series, node-link diagram of direct acyclic graphs (DAG) for topical transition over time), composing them into a single display in a meaningful and readable manner will be difficult. In searching the design space, we should also consider the dynamic relation among different types of entities (i.e., topics and opinion leader groups), as implied by the semantic relations among the temporal data (such as the competitiveness of each topic vs. the competitiveness attributed to each opinion leader group). To formulate explanatory hypotheses for the observed patterns (Q5, Q6), it is also crucial to provide a concise visual summary for the contents of the tweets, which can be dynamically queried and compared along temporal, social (i.e., different opinion leader groups) and topical dimensions.

5.3 Visual Encodings

In this section, we describe the visual encoding scheme and how each research question can be explored using the visual designs.

5.3.1 Timeline view

The timeline visualization provides an intuitive means of integrating the multiple types of temporal data (section 5.2). The theme of the design is to display the co-evolutionary relations between the topics and the opinion leader groups. A composite style is employed: the surge and decline of the competitiveness of the topics and the temporal varying salience are depicted using ThemeRiver, while the opinion leader groups are drawn as threads that are placed at the layers in the ThemeRiver given that they pose influence on the corresponding topic. These threads will switch to different layers as opinion leaders change their topical focus and began to play a major role in recruiting audiences for those topics as they set the public agenda. Figure 4 shows a timeline visualization based on the design.

The threads can appear bundled or separated as the opinion leaders' roles in influencing the public agenda change over time. The visual design employs the analogy of the diverging and converging behavior of social entities. Bundling of the lines indicates commonality, whereas separation indicates the contrary. The metaphor has been used in numerous visual designs, where the behavior can be the affiliation to social communities [34] or its presence in different events [39]. In this case, it is the role that the opinion leaders play in the agenda-setting process.

The composite visual design which integrates various types of temporal data will make identifying temporal correlation patterns more efficient than when data are depicted separately [5]. Several patterns can be identified from the visual design, such as: 1) the correlation between the topical transition of one or more opinion leader groups and the increasing or decreasing competitiveness of the topics; 2) temporal salient correlation patterns such as the competitiveness of a topic is mostly related to a constant set of opinion leader groups for a long duration; 3) two or more opinion leader groups contribute to the competitiveness of the same topic at a certain time but diverges afterwards (Q1, Q2, Q3). We describe the details of the visual encoding scheme as follows:

Layers in the ThemeRiver Each layer corresponds to a topic that constantly engages the attention of the public. The **height** of each layer is proportional to the competitiveness of the topic, which is the most important dimension in the analysis. Given that the measure of competitiveness is directly comparable in multiple topics, the **accumulated height** can reveal the overall intensiveness of the competition among them. **Color intensity** encodes the percentage of public engagement (i.e., saliency of the topic). Topics that engage more public attention will be assigned higher values.

Line segments During each time interval, a line segment is drawn on the layers if the recruitment effect of an opinion leader group on the corresponding topics (defined in section 4.2) is upon a threshold on adjacent timeframes, thus playing a (relatively stable) role in drawing the attention of the public from other topics as they set the public agenda. The **width** of the lines is proportional to the competitiveness contributed by each opinion leader group to the topic. Thicker line indicates a more important role. **Categorical colors** differentiate the opinion leader groups.

Transition lines The transition lines which connect the threads on different layers encode topical transition, which could aid in the explanation of the dynamically changing roles that the opinion leaders play in the agenda-setting process.

How to measure the amount of topic transition is one problem that has to be addressed. We use a *soft matching* approach, inspired by [15] which uses transition matrix to measure the flow of people among evolutionary social communities. Our method estimates the average trend of topical focus transition for a group of people at neighborhood time intervals between every pair of topics. Through the following least squares formulation the average transition, described as a $k \times k$ matrix (i.e., A in Equation 3), can be estimated:

$$\begin{gathered} \min \sum_{l} \omega_l ||m_l^{t-1}A - m_l^t||^2 \\ \text{subject to}: \sum_{j=1}^{k} a_{ij} = 1 \text{ and } a_{ij} \geq 0 \end{gathered} \tag{3}$$

In equation 3, m_l^{t-1} is a vector with values describing the engagement of an opinion leader l in the k topics (i.e., the percentage of tweets posted on each topic) at time $t-1$. In matrix A, each entry a_{ij} describes the trend of topic transition (i.e., "permutation") from topic i to j between adjacent time interval $t-1$ and t for a group of opinion leaders. If a_{ii} is close to 1, the engagement of the opinion leader group on topic i remains stable. The entries in A are estimated such that the sum of the squared error between $m_l^{t-1}A$ and m_l^t is minimized. a_{ij} should also satisfy the constraints presented in Equation 3 such that the result will be interpretable. ω_l is the weight related to opinion leader l, for which the number of tweets is used in the current implementation.

The transition matrix is estimated for each opinion leader group at adjacent time intervals with a standard linear constrained quadratic programming procedure [6]. The entry a_{ij} denotes the amount of topical focus transition from topic i to topic j for a group of people. If the amount of transition between $t-1$ and t is larger than a specified threshold, and in the meanwhile the opinion leader group's influence on i exists on $t-1$ but switch to j on time t, the entry will be drawn such that it could explain the change in the role of the opinion leader group. To encode the values of a_{ij}, we use dashed lines with varying **density of dots**. The density of the dots is proportional to the strength of the transition. The dotted line still preserves visual continuity while visually diminishes according to the Gestalt principle [42], indicating a more vague relation when the transition does not appear to be evident.

5.3.2 Radial view and word cloud

The radial view displays pairwise competition among topics in a selected time interval (Q4). This view extends the graphical notation originally used in the research domain [50] to include additional competing topics. Figure 4(c) shows an example of the radial view displaying pairwise competitiveness. The topics are arranged as nodes on a circle, whereas pairwise competition among the topics is displayed as edges routed within the circle. The width of the edge indicates the strength of interaction. The direction of the color gradient indicates the direction of recruitment effect. The size of each node is proportional to the competitiveness of the corresponding topic. Color intensity encodes the saliency, which is the same in the timeline view for consistency.

Each node also provides information on the competitiveness contributed by each opinion leader group in the selected time interval by displaying a color-coded pie chart upon user selection. In other words, the pie chart displays the proportion of the competitiveness contributions of different opinion leader groups to the corresponding topic. The radial view also serves as a widget for exploring tweet contents. A word cloud is displayed as user interacts with the nodes and pie charts to query on multiple facets, including the opinion leader groups, and the topics. The user can seek for the underlying causes for the patterns observed in the timeline view (e.g. topic transition, increase in competitiveness) with the widget. The radial view and the word cloud are displayed on demand as users make selections on the timeline view. Multiple word clouds can be pinned on the display to allow efficient comparison across multiple dimensions thus enabling the user to explore Q5 and Q6.

The size of each word in the word cloud is proportional to the *tf-idf* (term-frequency inverse-document-frequency) measure [36]. The method takes the raw frequency of the terms in the selected collection of Tweets and down-weight them by the total amount of occurrences of the terms in the documents, thus more unique and salient terms will be given more weight. The temporal dimension is considered by treating the tweets posted in each time range (days or weeks) as a document when computing the *idf*. Words that only appear during specific time intervals are assigned higher importance. It would make comparative analysis across the temporal dimension easier. To facilitate the analysis of the correlation of keywords to the observed recruitment effects, we draw a sparkline [22] denoting the temporal variation of the keywords' occurrences for the entire time range under analysis upon user selection.

5.4 Interactions

Various interactive features are provided in the system to support investigative analysis and the exploration of the research questions (Q5,Q6).

Detail-on-demand There are multiple levels of detail available to the users. The word cloud in the radial view gives a summary of the textual contents of the tweets in a selected time range and facilitates the identification of significant topics. In the word cloud, a sparkline denoting the temporal variation of the keywords occurrences will also show up upon user selection. A list of raw tweet records us also provided to the users such that they can gain contextual understanding for the keywords in the wordcloud by reading the tweets.

Comparative analysis In the radial view, the user could select on the nodes correspond to the topics and the segments on the pie charts correspond to the opinion leader groups. The word clouds would then provide a visual summary of the textual contents of tweets filtered from the two dimensions. As multiple word clouds can be simultaneously brought forth, side-by-side comparison is possible.

6 IMPLEMENTATION

In this section, we describe the technical details in implementing the components of the visual analysis framework as illustrated in section 3.

Data analysis. By submitting keywords and time range queries to the search engine and by recording the number of matches, the time variant saliency of the topics (i.e. $m_{i,g}$ and p_i in Equation 2) can be obtained. Given the m and p time series, we perform an ordinary least squares estimation using the *vars*[2] package in R which generates the estimated coefficients (i.e., $\beta_{i,j,g}$) in Equation 2 and the R^2, which describes the overall fitness of the model. A stepwise regression is performed thereafter to obtain the effect for each recruitment term (i.e., $sr_{i,j,g}^2$ for $m_{i,g}p_j$), which are aggregated to obtain the measures on the competitiveness of each topic and the recruitment effect of each opinion leader group with respect to the topics. To obtain time varying measures, we use a sliding window with a fixed number of time points. The data for estimating topic transition is also obtained through keyword queries. Given the corresponding recruitment effect of each opinion leader group, and the topic transition trend at each time point, a DAG structure that encodes the varying roles and the topical transition of the opinion leaders is derived, which are then displayed with the timeline view. The word cloud is constructed by fetching information from the search engine upon user interaction.

Timeline view. Two steps are used to create the timeline view: to order the layers and threads and to derive their exact geometry.

To order the layers, we adopt the approach mentioned in the paper [7], which places the layer with the least amount of change (variance in

[2]`http://cran.r-project.org/web/packages/vars/`

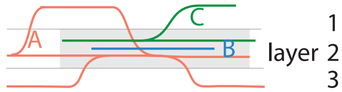

Fig. 5. The gray rectangle marked out a segment in layer 2. The order of insertion in the greedy scheme will be (A, C, B), as in descending order of the number of transition lines. After the rearrangement, the order of the threads from top to bottom will be (C, B, A).

competitiveness) at the center, and gradually add the layers with larger variance from inside out. After fixing the order of the layers, the order of the threads which co-exist in the same layer needs to be decided. Here a simple strategy can be used to reduce the resulted crossings and generate a more aesthetic layout. The method is illustrated in Figure 5. Here we define a continuous time interval when multiple threads co-exist in a layer as a *segment*. Within each segment, the order of the threads is decided with a greedy scheme by adding the threads one-by-one. As the crossings would be caused by the transition lines going to other layers, the threads with more transition lines are added first. Each time a new thread is added to the ordering, the place of insertion is determined through minimizing the resulted new crossings.

After the orderings have been decided, the exact geometry of the layers and the threads can be derived. The baseline (i.e., the bottom of the lowest layer) in the ThemeRiver is computed by minimizing the wiggles [7]. The threads are drawn close to the midpoints of each layer thus they appear bundled when multiple opinion leader groups exert influence on the same topics.

7 Evaluation

To evaluate the effectiveness of the model and to demonstrate the analytical and visualization techniques, we apply them on two data sets retrieved from Twitter and discuss our findings. We also demonstrated the prototype visualization system to domain experts who are the coauthors of this paper and gathered feedback on the effectiveness of the system.

7.1 Data Preparation

We use two Twitter data sets for our study: the 2012 US presidential election and the Occupy Wall Street movement (OWS). Both data sets were obtained by retrieving tweets with related keywords or hashtags such as "election2012" and "occupywallst". The time duration for retrieving the US election data was from May 01, 2012 to November 20, 2012, which covered the major events related to presidential election. A total of 89,174,308 tweets were obtained for the election data. The OWS data set contained 3,201,119 tweets posted from September 17, 2011 to November 25, 2011.

We worked closely with domain experts to determine the categories of opinion leaders and to identify the most important topics in the two data sets. The opinion leaders were detected through the number of retweets. For the election data set, the 200 most retweeted users, whose tweets accounted for 0.75% of the total tweets, were regarded as the opinion leaders in our study. The opinion leaders were manually classified by collaborating with domain experts into three groups, namely, *political figures*, *media*, and *grassroots* for the comparative study of their influences on the public agenda. Our collaborators also identified six general topics, namely, *welfare/society*, *defense/international issues*, *economy*, *election (general)*, *election (horse race)*, *law/social relations* considering their relevance to agenda-setting research [27] and to the dataset. Keywords related to the topics were collected through an iterative process which attempted to balance the percentage of tweets related to any of the six topics at different times while maximizing the overall coverage of the tweets. This process is suitable for studying the competition effect among the topics, as suggested by our collaborators. The final set of keywords covered 51.2% of the total tweets (averaged over all the four-hour intervals) with $std = 0.08$. The saliency of the topics on the public agenda was computed by dividing the number of tweets related to a particular topic by the total number of tweets posted by all users at a certain time. The agenda of each group of opinion leaders was obtained in a similar manner by counting only

Table 1. Evaluation of the model against three common measures in time series data analysis shows that the model is highly effective and robust. The table shows the average and the standard deviation (in parentheses) of the measures when applying a moving window estimation for the *2012 presidential election data*.

	Economy	Horse Race	Election General
R^2	0.98 *(0.01)*	0.98 *(0.01)*	0.99 *(0.00)*
$se_{\hat{y}}$	0.02 *(0.006)*	0.02 *(0.008)*	0.02 *(0.006)*
DW - d	2.13 *(0.18)*	2.17 *(0.18)*	2.14 *(0.16)*
	Defense / International	Law / Social Relations	Welfare & Society
R^2	0.97 *(0.02)*	0.96 *(0.03)*	0.95 *(0.04)*
$se_{\hat{y}}$	0.02 *(0.006)*	0.01 *(0.008)*	0.02 *(0.012)*
DW - d	2.18 *(0.18)*	2.11 *(0.17)*	2.11 *(0.18)*

the tweets posted by the corresponding opinion leaders. The definition of saliency of the topics has been confirmed by our collaborators.

For the OWS data, we identified three topics, namely, *protest activity (general)*, *corruption*, and *income inequality*. Then, we collected the keywords by following the same procedure used in the election data. We identified 100 opinion leaders whose number of tweets covered 2% of the total tweets. The opinion leaders were classified into three groups, namely, *protest accounts*, *media* and *grassroots*. The final set of keywords covered 44.8% (average) of the total tweets with $std = 0.06$.

The data which served as input for estimating the regression system in Equation 2 contained the agenda of the opinion leader groups and the public for each four-hour interval. A moving window estimation was performed, wherein the size of each time window spanned two weeks and contained 84 time points. Estimation was performed by moving the time window for one week for each estimation. The competitiveness of each topic and the influence of the opinion leaders were determined for each estimate.

7.2 Model Evaluation

We applied three measures commonly used in time-series analysis to evaluate the extended model using the election data, namely, the overall goodness of fit (R^2) of the regression model, the standard error of the estimates ($se_{\hat{y}}$), and the presence of autocorrelation in the residuals (Durbin-Watson d[12]). R^2 indicates the explanatory power of the model. As shown in Table 1, the mean value of R^2, averaged from the 1,241 date points of the time series for the equations of each topic, ranged from 0.95 (on *welfare / society*) to 0.99 (on *election general*), thus suggesting that more than 90% of the fluctuations in public attention on the six topics were explained by the model, with only 1% to 5% unaccounted for. $se_{\hat{y}}$ describes the predictive power of the model. Table 1 shows that the mean value of se_y varied in a narrow range within 0.01 to 0.02 on a scale of 0 to 1, thus suggesting that the errors of the predicted dependent variable fall within the narrow range of 2% to 4% at the 95% confidence level. DW-d determined the presence of autocorrelation between adjacent residuals of the model, with d ranging from 0 (perfectly positive autocorrelation) to 4 (perfectly negative autocorrelation). Table 1 shows that the d values were all close to 2 (that is, the absence of autocorrelation), thus suggesting that the residuals of the equation system were essentially white noise. The experiment indicates that the model appears to be highly effective and robust.

7.3 The 2012 US Presidential Election

In the first case study, we analyzed the competition among the topics that were either directly related to the 2012 US presidential election, or those that could trigger much public interest (such as the economy) in the course of the campaign period.

Figure 4 demonstrates a number of interesting patterns identified using the timeline view of the visual analysis system. Several preliminary explanations could be formed. The timeline visualization indicates that for most of the time, the *political figures* group played a prominent role in attracting public attention to the topic *election (general)* in setting the agenda of the discussion on Twitter. During the time interval around August 28 and September 4, the threads are relatively thicker, indicating a stronger recruitment effect. The potential cause for the observed effects was likely to be the occurrence of some important events related to the topic. Therefore, we explored by examining the textual contents of the tweets during that time interval posted by the *political figures* on the topic *election (general)* with the word cloud. In the word cloud, the keyword "convention" was the most salient among all. Hovering over the keyword, a sparkline summarizing the temporal variation for the occurrence of the keyword "convention" was displayed. It could be identified that the occurrence of the keyword rose to its peak in the selected time interval (i.e., around September 4th). By examining the tweets that contains the keyword "convention", we hypothesized that the recruitment effect of the *political figures* with respect to the topic *election (general)* was very likely to be related to the events *Republican and democratic national convention*, both took place during that time.

During the time interval around October 2, the *grassroots* started to play a major role in recruiting public attention to the topic *election (general)* as they set the public agenda. In the meanwhile, it started to affect multiple other topics including *economy*, *welfare/society* and *defense/international issues*, which was very likely to be caused by the *presidential debates* that took place around that time, when various issues were brought to discussion. In the figure, we could also observe a stronger transition to the topic *welfare/society* than to *defense/international issues* as indicated by the two transition lines ***a*** and ***b***. We speculated that since the main topic of the first TV debate was on domestic issues, there could be relatively less topic focus transferring to *defense/international issues*. The figure also depicts the intensive competition among the topics during the presidential debates. The *grassroots* and the *media* were the two driving forces in attracting public attention for multiple topics. The radial view ***c*** in Figure 4 demonstrates that the topic *law/social relations* was the "victim" of some other topics around September 25.

Figure 6 illustrates another observation. The *media* exerted a long period of influence on the topic of *election (horse race)*. However, when the content of the discussion was displayed, we determined that the cause for the observed recruitment effect was actually different. Around July 10, a trending discussion on Bain Capital, a company of which candidate Mitt Romney is a cofounder, was observed. After we examined the tweets posted by the *media*, we found that it was likely related to the news reports on the inconsistency between Mitt Romneys statement on the duration he served the company and what have been filed.

Figure 1 also illustrates different causes for the recruitment effect. From the timeline view, it could be identified that the *grassroots* and the *media* played the major roles in recruiting public attention to the topic *welfare/society* around June 5 and 12. Although around June 5, the trending discussion was on "solyndra", while around June 12, the keywords "teachers" and "firefighters" gained more significance.

Figure 7 illustrates how transition lines could hint on the switch in topical focus. It could be observed that the topical focus of the *media* transfers from multiple other topics to *law/social relations* around July 24. The keyword "gun" became increasingly important. After examining the tweets, we found that this was possibly caused by the Orlando gun shooting event, which raised a lot of discussion on gun laws.

7.4 Occupy Wall Street

Figure 8 illustrates a number of findings when we used the system to study the OWS data. We observed that the overall competitiveness of the three topics remained steady during the first two weeks of October, then increased rapidly after October 15, reached its peak around October 22, and gradually decreased. Such results aroused our interest in finding the reason for the sudden increase in overall competitiveness. We used our radial graph and other detailed views to dig deeper into the data, and discovered that numerous tweets referred to the support of President Obama to the protesters, such as "the White House issued a statement saying Obama is working for the interests of the 99%" on

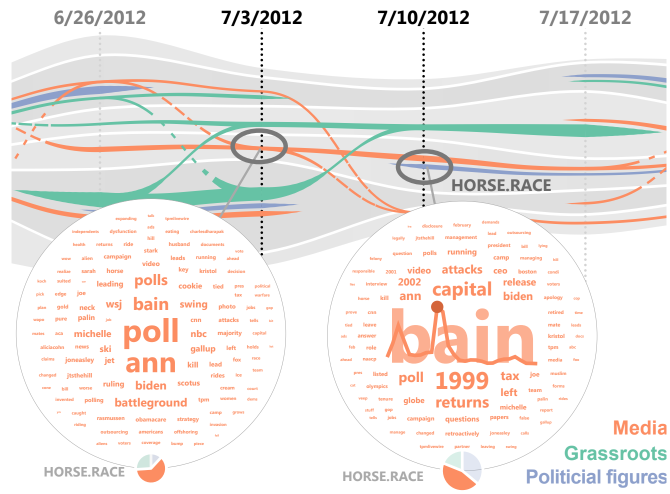

Fig. 6. A long time of influence exerted by the *media* on the topic *election (horse race)*, although with very different trending keywords.

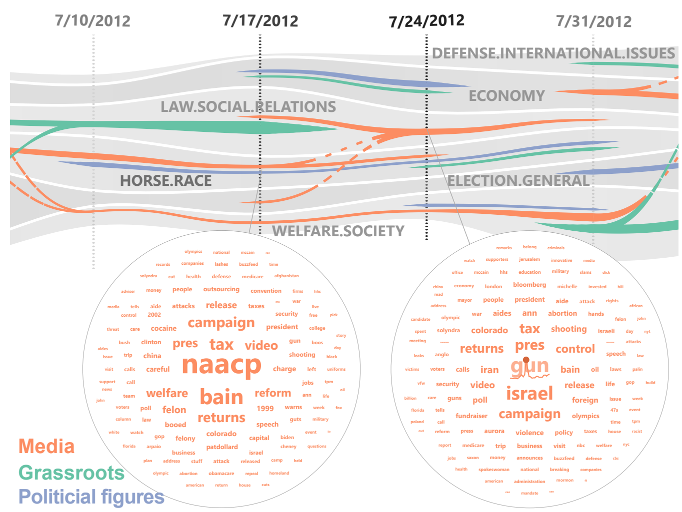

Fig. 7. Transition of topical focus of the *media* from multiple topics to *law / social relations* around July 24th. The keyword "gun" had an increasing importance when the word clouds based on all the tweets posted by the *media* were compared before and after the transition.

October 16. Thus, the support of President Obama could have encouraged the OWS movement and attracted public attention, which directly led to the increase in the overall competitiveness of all main topics.

We could clearly see that protest accounts played a major role in driving the public attention at all times. We examined the contents posted by the accounts of the protesters (bottom right word cloud in Figure 8), which featured the keywords "photo" and "posted", indicating the active involvement of these accounts in spreading the news and reporting the most recent protest activities. Another interesting pattern identified from Figure 8 was that "protest accounts" and "grassroots" dominated the discussion at the beginning. After a certain period, particularly after President Obamas statement of support, media groups began to play a more important role in leading the discussion.

Among the three topics, protest activity was the most salient because its corresponding layer in ThemeRiver has the darkest color. At the beginning of October, the grassroots were also quite active in protest activities. After examining the word cloud, we found that the keywords "arrested", "nypd", "police", "brooklyn", and "bridge" were among the most frequent. We further checked the Tweets and news reports around that period, and found that the observed recruitment effect was related to the arrest of protesters at Brooklyn Bridge on

October 2, which attracted much public interest during that time.

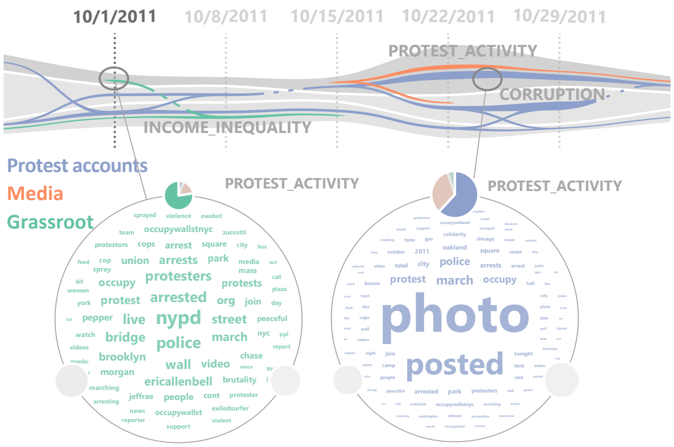

Fig. 8. Exploration of the Occupy Wall Street movement data. The *protest accounts* played the major role in recruiting public attention.

7.5 Expert Review

To evaluate the effectiveness of the visual analysis system, we interviewed two domain experts from two universities, A and B, regarding agenda-setting research and media theory, respectively. We first explained the visual encoding scheme used in the system and demonstrated its interactive features. We also demonstrated the patterns found in the case study. The initial feedback is summarized as follows.

Visual design. The domain experts were impressed with the visual design. Expert *A* commented "It is quite powerful and help in my research. The visualization can help reveal the complex, dynamic relationships among topics and opinion leaders visually. The visual representation appears intuitive to me". He further highlighted the usefulness and intuitiveness of the transition lines for connecting opinion leaders in his analysis. Expert *B* also appreciated our timeline visualization system and acknowledged its usefulness and effectiveness. He further added "The radial view is easy to understand as it is commonly used in my field to represent the pairwise competition". He also pointed out that the tweet list will be helpful for in-depth analysis of certain patterns identified from other visualizations.

Interactive features. The interactive features were also well received . They both appreciated using our system. Expert *A* acknowledged the intuitiveness of the interactions provided by the system. He said "I like interacting with the data and seeing the results immediately. The interactions are very smooth". He also considered the interactions to be engaging. Expert *B* particularly liked the feature of generating a radial visualization by clicking any part of the timeline visualization. He also believed that the interactive comparative analysis supported by our system is useful and intuitive.

Improvements. The experts recommended several potential improvements, including adding more legends and textual explanations to the quantities encoded in the system. They also suggested that the visualization system be released to the public via a Web-based environment. Expert *A* commented "While the current system looks nice and comprehensive, it may not fully meet my requirements. I would like to customize the system if possible. For example, I might be only interested in fewer topics and so other topics could be removed from the visualization". Expert *B* also highlighted the idea that the system should start simple and hide unnecessary details. He suggested that the system should initially show basic information (such as the competitiveness of each topic) with ThemeRiver. The threads representing the opinion leader groups can be displayed upon user request. All suggested improvements were integrated in the current system.

8 Discussion

The case studies demonstrate the advantages of combining a quantitative competition model and interactive visualization techniques to discover interesting patterns. The model is capable of extracting structured relation information from massive unstructured tweets, providing analysts with a new method for gleaning insight. Nevertheless, the modeled information is dynamic and poses an obstacle to the understanding. Our interactive visual analysis technique can convey information in an intuitive manner, mainly through timeline visualization. This study mainly analyzes the dynamic relationship between opinion leaders and various topics in the agenda-setting process. However, the major component of the visual design (that is, the timeline view) can also be applied to other scenarios when a time-varying relation exists between two different entities. For example, the relationship between social network users and time-evolving communities, wherein each user switches their affiliation with the communities.

In our study, six broad topics are extracted from millions of tweets on Twitter for the presidential election data. Considering the relative proportion of the topics under the analysis, we measure the saliency of the topics as a normalized score (i.e., the percentage of public engagement in those topics) rather than a raw count (i.e., number of tweets about those topics). The normalization of the data structurally determines the agenda-setting process to be a competition process. Therefore, it is quite reasonable for us to focus on the competitive relationship among topics only in our study. Although competition has been a popular framework in the study of information diffusion [4, 45, 48, 50], it is worth noting here that there might be other forms of relationships such as cooperation between topics [31]. To detect other forms of relationships, a feasible direction is to zoom into these six broad topics identified in the study and decompose them into several sub-topics. When more subtle topics included in the study, it is more likely to find different forms of relationships among the topics.In the current study we mainly analyze the competitive relationship. In the future, we plan to extend the model to include other forms of relationships.

Our research on the phenomenon of topic competition in agenda-setting is still in progress. Two weaknesses in this work need to be addressed. First, data is still far from perfect and may lead to biased conclusions. Second, opinion leader groups and topics are defined and provided by our domain experts for both data sets. To apply the model to other data sets, we also need to define topics and opinion leader groups, which is time-consuming and tedious. It is certainly possible to classify the tweets and extract opinion leaders automatically or semi-automatically using data mining approaches. However, automatic tweet classification and opinion leader detection is beyond the focus of this work. We will study this problem and integrate the data mining techniques into the current system to automate the data collection process.

9 Conclusion

This paper describes a visualization system for facilitating the analysis of the competition effect among multiple topics in agenda-setting on social media. We introduce a model for characterizing the dynamics of topic competition and for estimating the influence of different opinion leaders. The system employs a timeline design which integrates stacked graph and storyline visualizations. The system allows analysts to interactively and visually trace and analyze the estimated dynamic relationships among competing topics, as well as the relationships between opinion leaders and the topics over time. A set of detailed views, such as radial graph and word cloud, is provided for in-depth analysis of the pattern discovered from timeline visualization. In the future, we will test our models and techniques on more data sets. Our system is currently off-line. An interesting direction will be extending it to support real-time Twitter data streaming. We plan to conduct a formal user study to assess the effectiveness of the visual design and the utility of the system.

Acknowledgments

We would like to thank Professor Lu Wei from Zhejiang University for suggesting the chance of interdisciplinary collaboration, the anonymous reviewers for their valuable and constructive comments, and Fangzhao Wu for his help. This work is partially supported by a grant from MSRA and RGC GRF 618313.

REFERENCES

[1] Project cascade. http://nytlabs.com/projects/cascade.html, Sept. 2011.

[2] Revisit. http://moritz.stefaner.eu/projects/revisit/, Sept. 2012.

[3] D. Archambault, D. Greene, P. Cunningham, and N. Hurley. Theme-Crowds: Multiresolution summaries of twitter usage. In *Proceedings of the Workshop on Search and Mining User-generated Contents*, pages 1–20, 2011.

[4] S. Asur, B. A. Huberman, G. Szabo, and C. Wang. Trends in social media: Persistence and decay. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 434–437, 2011.

[5] M. Q. W. Baldonado, A. Woodruff, and A. Kuchinsky. Guidelines for using multiple views in information visualization. In *Proceedings of Advanced Visual Interfaces*, pages 110–119, 2000.

[6] S. Boyd and L. Vandenberghe. *Convex Optimization*. Berichte über verteilte messysteme. Cambridge University Press, 2004.

[7] L. Byron and M. Wattenberg. Stacked graphs - geometry & aesthetics. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1245–1252, 2008.

[8] N. Cao, Y.-R. Lin, X. Sun, D. Lazer, S. Liu, and H. Qu. Whisper: Tracing the spatiotemporal process of information diffusion in real time. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2649–2658, 2012.

[9] 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 Transactions on Visualization and Computer Graphics*, 17(12):2412–2421, 2011.

[10] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, pages 115–122, 2010.

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

[12] J. Durbin and G. S. Watson. Testing for serial correlation in least squares regression. *Biometrika*, 37(3/4):409–428, 1950.

[13] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, volume 1. Cambridge University Press, 2010.

[14] E. R. Gansner, Y. Hu, and S. C. North. Visualizing streaming text data with dynamic graphs and maps. In *Proceedings of Graph Drawing*, pages 439–450, 2012.

[15] M. Gupta, J. Gao, Y. Sun, and J. Han. Integrating community matching and outlier detection for mining evolutionary community outliers. In *Proceedings of the ACM SIGKDD international conference on Knowledgediscovery and data mining*, pages 859–867, 2012.

[16] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. ThemeRiver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002.

[17] M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K.-L. Ma. Breaking news on twitter. In *ACM CHI*, pages 2751–2754, 2012.

[18] J. Johansson, P. Ljung, and M. Cooper. Depth cues and density in temporal parallel coordinates. In *Proceedings of EuroVis*, pages 35–42, 2007.

[19] N. W. Kim, S. K. Card, and J. Heer. Tracing genealogical data with timenets. In *Proceedings of Advanced Visual Interfaces*, pages 241–248, 2010.

[20] R. Kosara, F. Bendix, and H. Hauser. Time histograms for large, time-dependent data. In *Proceedings of EuroVis*, pages 45–54, 2004.

[21] S. Laxman and P. Sastry. A survey of temporal data mining. *Sadhana*, 31(2):173–198, 2006.

[22] B. Lee, N. H. Riche, A. K. Karlson, and M. S. T. Carpendale. Sparkclouds: Visualizing trends in tag clouds. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1182–1189, 2010.

[23] S. Liu, Y. Wu, E. Wei, M. Liu, and Y. Liu. Storyflow: Tracking the evolution of stories. *IEEE Transactions on Visualization and Computer Graphics*, 19(12), 2013.

[24] 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 Transactions on Intelligent Systems and Technology*, 3(2):25:1–25:28, Feb. 2012.

[25] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. A. Keim. EventRiver: Visually exploring text collections with temporal references. *IEEE Transactions on Visualization and Computer Graphics*, 18(1):93–105, 2012.

[26] A. Marcus, M. Bernstein, O. Badar, D. Karger, S. Madden, and R. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *ACM CHI*, pages 227–236, 2011.

[27] M. Mccombs and J. Zhu. Capacity, diversity, and volatility of the public agenda: Trends from 1954 to 1994. *Public Opinion Quarterly*, 59(4):495–525, 1995.

[28] M. E. McCombs and D. L. SHAW. The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2):176–187, 1972.

[29] W. Muller and H. Schumann. Visualization methods for time-dependent data - an overview. In *Proceedings of the Simulation Conference*, volume 1, pages 737–745 Vol.1, 2003.

[30] R. Munroe. Xkcd #657: Movie narrative charts. http://xkcd.com/657, December 2009.

[31] S. Myers and J. Leskovec. Clash of the contagions: Cooperation and competition in information diffusion. In *Proceedings of IEEE International Conference on Data Mining*, pages 539 –548, 2012.

[32] M. Ogawa and K.-L. Ma. Software evolution storylines. In *Proceedings of the international symposium on Software visualization*, pages 35–42, 2010.

[33] V. Ogievetsky. Plotweaver. http://ogievetsky.com/PlotWeaver/.

[34] K. Reda, C. Tantipathananandh, A. E. Johnson, J. Leigh, and T. Y. Berger-Wolf. Visualizing the evolution of community structures in dynamic social networks. *Computer Graphics Forum*, 30(3):1061–1070, 2011.

[35] S. Rose, S. Butner, W. Cowley, M. Gregory, and J. Walker. Describing story evolution from dynamic information streams. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, pages 99 –106, 2009.

[36] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, Nov. 1975.

[37] C. Shi, W. Cui, S. Liu, P. Xu, W. Chen, and H. Qu. RankExplorer: Visualization of ranking changes in large time series data. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2669–2678, 2012.

[38] L. Shi, F. Wei, S. Liu, L. Tan, X. Lian, and M. X. Zhou. Understanding text corpora with multiple facets. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, pages 99–106, 2010.

[39] Y. Tanahashi and K.-L. Ma. Design considerations for optimizing storyline visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2679–2688, 2012.

[40] J. van Wijk and E. Van Selow. Cluster and calendar based visualization of time series data. In *Proceedings of IEEE Symposium on Information Visualization*, pages 4–9, 140, 1999.

[41] F. Viégas, M. Wattenberg, J. Hebert, G. Borggaard, A. Cichowlas, J. Feinberg, J. Orwant, and C. Wren. Google+ ripples: A native visualization of information flow. In *Proceedings of the international conference on World Wide Web*, pages 1389–1398, 2013.

[42] C. Ware. *Information Visualization: Perception for Design (Interactive Technologies)*. Morgan Kaufmann, 1st edition, Feb. 2000.

[43] M. Wattenberg and J. Kriss. Designing for social data analysis. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):549–557, 2006.

[44] M. Weber, M. Alexa, and W. Muller. Visualizing time-series on spirals. In *Proceedings of IEEE Symposium on Information Visualization*, pages 7–13, 2001.

[45] L. Weng, a. Flammini, a. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Scientific reports*, 2:335, Jan. 2012.

[46] K. Wongsuphasawat and D. Gotz. Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2659–2668, 2012.

[47] D. K. Wright and M. D. Hinson. How blogs and social media are changing public relations and the way it is practiced. *Public Relations Journal*, 2(2):1–21, 2008.

[48] F. Wu and B. A. Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601, 2007.

[49] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In *Proceedings of the international conference on World Wide Web*, pages 705–714, 2011.

[50] J. Zhu. Issue competition and attention distraction: A zero-sum theory of agenda-setting. *Journalism & Mass Communication Quarterly*, 69(4):825–836, 1992.