Visual analysis of retweeting propagation network in a microblogging platform

Quan Li, Huamin Qu, Li Chen, Robert Wang, Junhai Yong, Detan Si

Visual Analysis of Retweeting Propagation Network in a Microblogging Platform

Quan Li
School of Software
Tsinghua University
whuisslq@163.com

Li Chen
School of Software
Tsinghua University
chenlee@tsinghua.edu.cn

Junhai Yong
School of Software
Tsinghua University
yongjh@tsinghua.edu.cn

Huamin Qu
Computer Science and Engineering, Hong Kong University of Science and Technology
huamin@cse.ust.hk

Robert Wang
Information System, Sina Technology, (China)Co., Ltd
robert1@staff.sina.com.cn

Detan Si
Information System, Sina Technology, (China)Co., Ltd
sidetan@staff.sina.com.cn

ABSTRACT

As a novel type of real-time social networking service, microblogging has already become ubiquitous and an irreplaceable tool. Tracking in the pulse of retweeting propagation is important and meaningful. In this paper, we investigate how information propagation in a specific microblogging platform evolves to identify relevant patterns and understand dynamic attributes of information propagation and the underlying sociological motivations. More specifically, based on the node-link diagram, we propose three efficient strategies to map the multiple attributes of information propagation graph to appropriate visual elements. For revealing the dynamic attributes, we propose two models: the depth-varying and the time-varying parallel data model to illustrate the temporal evolution efficiently. We also present a novel method by combining the traditional scatter plot with Hough transformation to represent the distribution of propagation instances and trace the propagation speeds. We integrate our methods to a visual mining tool and develop several interactive features. We demonstrate how our approaches improve the understanding of the propagation graph from a visual perspective by employing propagation datasets collected from Sina Weibo, the largest microblogging service provider in mainland China. Meanwhile, this visual mining tool has been evaluated by data analysts and successfully used in Sina Corporation as a helpful assistant to them.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: General; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Graphical user interfaces (GUI)*; E.1 [**Data**]: Graphs and Networks

General Terms

Algorithm

Keywords

retweeting propagation, dynamic graph visualization, interaction, visual mining, micro-blogging

1. INTRODUCTION

As a type of social networking service, microblogging, such as Twitter, Sina Weibo has already become ubiquitous in Web 2.0 era. Tracking in the pulse of retweeting propagation is important. For corporations, it enables them to get feedback of user coverage and gain insight into how to improve and market better. For individuals, the abundance of information and opinions from diverse sources in microblogging platforms helps them tap into the wisdom of crowds, understand the retweeting behavior and aid in making more informed decisions [16]. Information propagation in social networks takes on different structures and scopes. However, little work has been done to reveal the coupled dynamics of both the structure and the information propagation. Most researchers have focused on the qualitative study or mathematical modeling of information propagations in social networks [10]. By contrast, in this paper, our objective is to track information propagation and mine the patterns from a visual perspective.

How to represent the graph of social network has been heavily researched [9]. The major challenge is the computing cost of graph drawing criteria. As the information propagation network involves multiple attributes and dynamic evolution, how to model the multi-attributes and map them efficiently to visual elements is a big challenge. A prevailing approach to deal with the dynamic attribute is to use animation [4][6]. However, animation mapping leads to high cognitive efforts due to human's limited short term memory. It is difficult to observe trends over longer periods or compare two non-subsequent time steps. Abrupt changes of the graph might furthermore destroy the viewer's mental image of the graph [2].

In this paper, we investigate how information propagation in a specific microblogging platform evolves and carry out

visual mining tasks on the social network graph. Our visualization tool provides several interactive features to explore dynamic graph data. We demonstrate how the above techniques and methods increase the understanding of the information propagation from a visual perspective in several application scenarios. Meanwhile, this system has been evaluated and used in the Weibo department of Sina Corporation. The major contributions of this paper are: (1): A comprehensive visual analysis system for mining retweeting propagation patterns in a microblogging platform; (2): Three novel visualization schemes, hierarchy-aware layouts, an accumulated filtering scheme and micro-aggregation, for multi-attribute propagation graphs; (3): Novel depth-varying and time-varying data models and visualization for scalable graph visualizations; (4): Case studies with real data to demonstrate the effectiveness of the system.

The remainder of this paper is organized as follows. We introduce the related backgrounds and existing work and then we describe our data and the multi-attribute graph. After that, we propose three novel visualization schemes to represent multiple social attributes regarding to specific business requirements. Then, we deal with the dynamic attributes. Following are several interactive features. Finally, two case studies are introduced to demonstrate the effectiveness of our system.

2. RELATED WORK

The related work falls into two categories: information propagation on social network and how to visualize the dynamic propagation patterns.

2.1 Information propagation on social network

Information propagation in social networks has been heavily studied in the past few years, especially with the rapid emergence of microblogging services, and has drawn considerable research attention in the area of data mining. In 2009, Sun et al. [19] investigated the propagation of Facebook's News Feed and created a social network based on users and fans of Facebook for analysis. Kwak et al. [15] used the relationships existed in the followers and fans in Twitter to construct a social network for analysis. Sakaki et al. [17] proposed a real-time event detection system using Twitter data. Diakopoulous and Shamma demonstrated the use of timeline analytics to explore the 2008 Presidential debates through Twitter sentiment [3]. Shamma et al. later carried out a similar analysis at the President Obama's Inauguration [18]. Gaffney performed a retrospective analysis of the recent Iran election, in which Twitter played a pivotal role in news reporting [8]. Jansen et al. found that tweet sentiment analysis could provide useful insights into product opinion, an idea leveraged for the sentiment visualization in TwitInfo [13]. Ho et al. focused on ways to measure how information is propagated and considered how to quantify a person's capability to disseminate ideas via a micro-blog and measure the extent of propagation of a concept in a micro-blog [11]. The above works use microblogging data for mathematical modeling or statistical analysis, without considering the dynamic and multiple attributes. Meanwhile, to the best of our knowledge, works on visualization and analysis for retweeting propagation are scarce, many of which use statistical representations, timeline, or social graphs to visualize the propagation graph.

2.2 Dynamic Propagation Visualization

The traditional node-link diagrams are effective to reveal propagation paths from a visual perspective. However, this graph layout cannot deal with massive graph data efficiently. The Force-Directed visualization layout model has been a widely adopted model, such as the classic Force-Directed algorithm [5], the KK algorithm [14], and the FR algorithm [7]. Researchers have done a lot of simulation to reduce the algorithmic complexity [1]. In our improved algorithm, we employ the traditional node-link diagram and the spring model. However, we optimize the initial position of each vertex before iterations.

As for the time dimension in dynamic graphs, traditional approaches use animation [6][21]. However, they require a heavy user's cognitive load. Burch, et al. proposed a novel method, mapping time to space and showing the subsequent graphs in a static diagram [2]. Our approach is inspired by their work. However, the graph model and mapping strategy in our work are quite different from theirs.

3. MULTI-ATTRIBUTE GRAPH MODEL

3.1 Difference Between Weibo and Twitter

Even though there existed several difference between the weibo and the western Twitter, we focus on the difference of replies, comments, and retweeting. For Twitter, the replies and comments appear independently in the feed, while on a weibo, they are listed under the entry, like a traditional blog. Replies are response to what someone else has written on Twitter whereas ReTweets are when you want to share what someone else has written with others so you copy and paste it as a Tweet update. The action of "retweeting", which is central in this paper, is a little bit different from that of the famous Twitter network. If you want to retweet one's tweet, you can directly retweet it or add your comments in addition to the original tweet and retweet them, provided that it does not exceed the maximum length of a tweet. Weibo supports the function of Reply and ReTweet simultaneously and you can add some comments in addition to the tweet. In this paper, we record the retweeting action, including retweeting without additional comments and with other comments.

3.2 Multi-Attribute Graphs

The propagation record consists of eight attributes. For all the records, the *rmid* (root message id) and *ruid* (root user id) are the same since all the retweeting behaviors aiming at a specific message posted by a specific user. In the propagation network, the message id and user id are unique. The *pmid* (parent message id) refers to the message id of the source vertex of the current retweeting occurrence and the *puid* (parent user id) represents the corresponding user id. The *cmid* (current message id) is the current message id of the retweeting occurrence and the *cuid* (current user id) is the corresponding user id. To map the multi-attribute information propagation graph into meaningful visual elements, we design a multi-attribute graph model. In this paper, we represent an information retweeting instance as a single node and an occurrence of retweeting behavior as a single edge. A user may retweet more than once, given three, for a single micro-blog, resulting in three vertices, representing three retweeting instances. Therefore, we consider our model as a Tree T with one vertex r as root node: $T = (V, E, r)$. The root node is the very beginning of propagation. Re-

lationships between vertices are expressed by T: Vertices sharing a common parent in T belong to the same "group". The *time* attribute represents the occurrence time of this retweeting behavior and the optional *gid* (group id) means the group id associated with this record. The retweeting propagation graph may become very large and is dynamic in nature, which pose special challenges for visualizations. In the next two sections, we will introduce some new methods we develop to address these two issues.

4. MULTI-ATTRIBUTE GRAPH

We present our methods to visualize large multi-attribute propagation graphs. First, we introduce a hierarchy-aware layout which can better reveal the retweeting relationship between nodes. Then we present a simplified layout to reduce the visual clutter for large data. Finally, a micro-aggregation scheme is proposed to visualize the groups in information propagation networks.

4.1 Hierarchy Aware Force-Directed Layout

In this section, we propose a Hierarchy Aware Force-Directed layout. We adopt the spring model and the Nbody algorithm and use a Quad-Tree data structure [1] to realize the Force-Directed layout. Each particle is stored in the tree structure, and corresponds to a leaf node in the tree. Each non-leaf node records the center coordinates of its children nodes and the sum of quality of all child nodes. When calculating the force of particles, we first traverse from the root node of the hierarchical tree. If the distance between the node and the particle exceeds a preset threshold value, we use the quality and the location of the node to calculate the force effect for particles. Otherwise, we continually traverse along the tree. More importantly, we change the initial position of each node according to its parent node's position in the hierarchy before each iteration, thus greatly decreasing the time complexity. The detailed framework is shown in the following steps.

(1) Instantiate the layout algorithm and add two sub-layout algorithms: Nbody algorithm and spring model.
(2) Calculate filtering condition depth: current time and depth value.
(3) Filter all vertices to invisible and set those satisfying the filtering conditions visible and put them into a set.
(4) Initial the vertex coordinates and positions in Set(V).
(5) Assign v's father node's position to v's current position.
(6) Iterate on the set and get all the visible edges.
(7) Calculate the force between the two vertex of each edge and initial the length of a spring to represent the property of the edge.
(8) Update nodes' position.

We use a dataset to compare the visualization results by adopting different algorithms. Firstly, we, respectively use the original Force-Directed layout algorithm and our improved method to show the graph with group clustering representation, the first time-step plot and the last time-step plot. We can clearly see our Hierarchy Aware Force-Directed layout is much clearer and has less visual clutter, especially when it reaches to the final time-step. Meanwhile, the group structures in the network are better preserved (Fig. 1).

4.2 Simplified Layout

Visualizing graphs in node-link diagrams may lead to visual clutter in the final display for large graphs even with an appropriate layout algorithm. In a microblogging retweeting propagation network, the most concerned part is the key nodes for retweeting. In our model, an edge represents a single retweeting behavior; therefore, nodes with higher out-degrees should be considered more significant according to the sociology and graph theory. However, it is not adequate to just consider the direct retweeting number (out-degree) that one single vertex brings in. The accumulated retweeting number that propagation paths starting from this vertex should be taken into account.

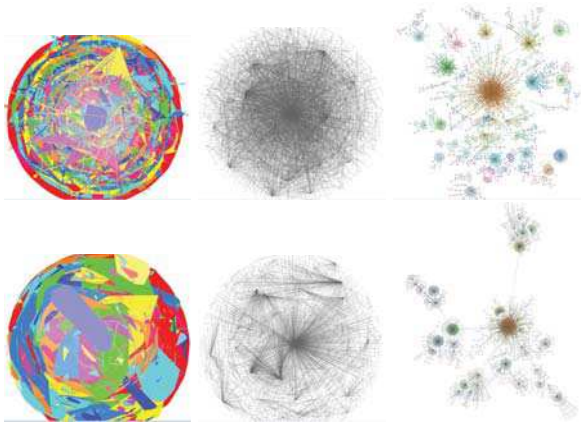

Figure 1: Visualization with the original Force-Directed layout algorithm (Top, From left to right: graph with group colors, first time-step, final time-step). Visualization with the Hierarchy Aware Force-Directed layout algorithm (Bottom, From left to right: graph with group colors, first time-step, final time-step). Colors represents different retweeting groups

4.2.1 Data Evaluation

We evaluate the simplified layout method based on one microblogging retweeting propagation dataset. It, which was firstly published by China Newsweek, covers 594 retweeting propagation instances. The longest path of propagation is 3 and the average path is 2.75. It starts at 2011-11-14 10:33:00 and ends at 2011-11-14 23:56:24, which lasts about 803 minutes. The original propagation graph and the simplified graph are shown in Fig. 2.

Figure 2: The original propagation graph and the simplified graphs (From top left to bottom right: the original propagation graph, the simplified graph with top retweeting number for each depth, the simplified graph with user's information, and the simplified graph with top three retweeting number for each depth)

When we visualize the top one team for each depth, we get the corresponding information of the most related users, as shown in Fig. 2. The number of fans and followers of "China Newsweek" are 1572, 079 and 1, 793, respectively. "Direct" means the number of nodes that directly retweet the message published by their superior parent in hierarchy. "Accumulation" refers to the number of all the nodes that directly or indirectly retweet the message. Of all the 8 nodes, the node of China Newsweek appears 4 times. One is functioned as the root node of this microblogging retweeting propagation network. The other three lay in the first depth of this network and act as retweeting nodes. We establish that the retweeting behavior of the root node itself plays a vital role in retweeting information.

4.2.2 *Propagation Analysis*

It is significant to trace the information flow and diffusion along the propagation path, especially those retweeting behavior and retweeting messages published by them. These key retweeting behaviors and the additional messages attached on the original message are essential for tracking the flow of sentiment through networks and emergence of polarization. Take this propagation dataset for example, at 11:21:35, China Newsweek retweeted its message published at 10:33:00. At 13:49:01, it retweeted again and the same went to the time of 15:20:19. We set the threshold accumulation value higher and visualizing the top three at each depth, and we get these messages: [10:33:00: kinship fall: who is "robbing" your donations?], [11:21:35 Dalian City: Parents hijack 80,000 donation; a two-year-old scalded child struggling in hospital], [13:49:01 Jiangsu Province: Teacher privately takes possession of donation; Students suffered and helpless], [14:20:58 Charity Collapse], [16:25:36 Every donation should be a hot potato on the hand], and [17:20:16 Charitable organization should not be bypassed].

We conclude that at this circumstance, each key retweeting behavior is another extension to the original topic/message. In other instances, the emergence of polarization may exists, since different polarizations arguing on a certain topic are likely to trigger large portion of web users' discussion and bring in new retweeting propagations. In this propagation network, "China Newsweek" is a public news media verified by Sina Weibo, therefore, it more likely retweets a summary of the related or similar topics, or the follow-up reports, which are likely to be proposed by single individuals.

4.3 Micro-aggregation

An important phenomenon for microblogging users is that the emergence of groups. The groups could be all the users having the same hobbies, or the same labels and all the corresponding topics gathered inside. When visualizing and mining these propagation records with "group" information, it is not intuitive to just plot them in the traditional node-link layout without giving more visual elements to help analysts better understand how the information flows within and between different groups. In this section, we propose a novel visual representation to visualize and analyze the microblogging propagation with aggregation information and evaluate with a real propagation dataset. This dataset covers 2000 instances. The number of micro-group is 38. We first collect all unique group ids and assign different color to each group and represent different groups as aggregation circles, then iterate all vertices of $Set(V)$ in $GraphVertex$, assign the vertex to the corresponding aggregation circle based on group id. The color of each vertex is based on its current user id (cui). We initially partition the screen according to the size of aggregation circles. For each aggregation group, we apply the Force-Directed layout algorithm to the vertices inside the group. The movement of the vertices is limited to the boundary of the aggregation group. For the connections between aggregation groups, we just link the vertices and cancel the force between these vertices.

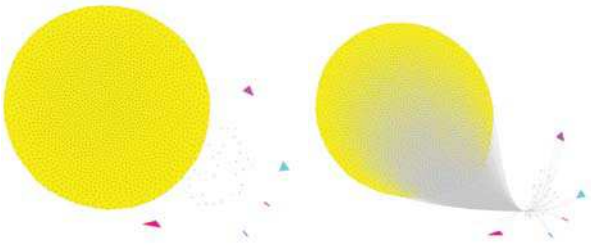

Figure 3: Visualization of micro-aggregation algorithm for the dataset with micro-aggregation information (From left to right: initial positions for each aggregation, inter/intra-connection between aggregations)

By observing Fig. 3, we find almost all the intra-connections happen from one specific aggregation group to another. By observing the color of vertices, we find that although there are 2000 vertices, they only correspond to 7 user ids, which means all the propagations are due to these 7 users. Among the 7 users, we find that one particular user is excessively active and occupies most of the propagations. By analyzing the information of this user, we find that this user has joined many different groups and has retweeted this microblogging message to all the groups. By observing the largest group, represented by yellow color, we find that this user had continually retweeted this message within this group, and no other users had published any messages, which means this user is continually refreshing the screen by retweeting the message inside the aggregation group. From this example, we can clearly see that micro-aggregation can significantly reduce visual clutter and reveal the groups in the propagation network.

5. SCALABLE DYNAMIC GRAPH VISUALIZATION

Another important and special feature of the propagation network is the temporal evolution. In this section, we propose a visualization method to use a single-image to explore dynamic and hierarchical propagation graphs. Two data models, the depth-varying data model and the time-varying data model have been proposed. In the depth-varying data model, we map each depth of the graph to a single axis on screen. The vertices belonged to each depth are arranged along the vertical axis according to their retweeting time. A line that connects the vertices starts from the left most side and ends at the right most side, as long as this propagation path has reached that depth. One link represents one propagation record. In the time-varying data model, we map each time step of the graph to a single axis on screen. The vertices belonged to each time step lies at certain posi-

tion according to their hierarchy in the graph. To give the distribution of clusters in graphs, we apply histogram view along the vertical axes.

5.1 Depth-varying Parallel Data Visualization

Based on the idea of parallel coordinates [12] and the depth-varying attribute in the propagation graph, we plot the depth-varying parallel data visualization. The result is as shown in Fig. 4.

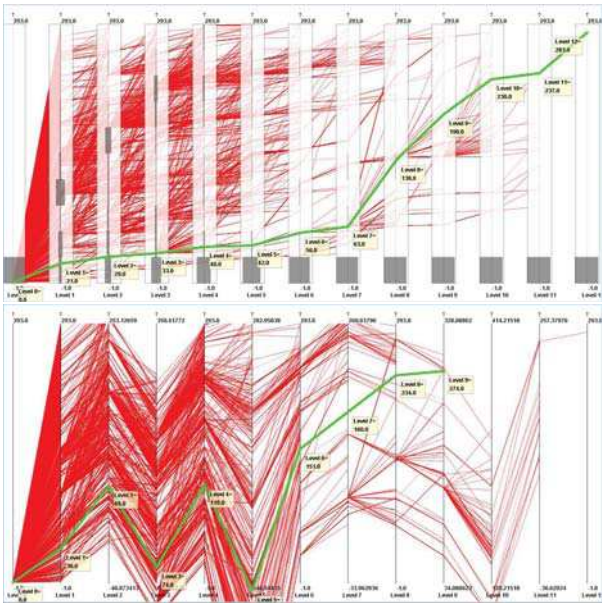

Figure 4: Depth-varying parallel data visualization (top) and interactions (bottom): dragging the depth axis up and down and highlight the propagation path

By a deep visual analysis with Fig. 4 (top), we can easily detect the longest propagation path, as the green line indicating. From the tooltip we can get more information along the longest propagation path, such as when the microblogging message arrives and the current depth. By showing the histograms of each axis, the distribution of vertices is quite intuitive. Meanwhile, we facilitate analysts by combining several interactive techniques and "Link + View" techniques. Fig. 4 (bottom) shows the related supporting interactive techniques: dragging the depth axis up and down to highlight the propagation path.

5.2 Time-varying Parallel Data Visualization

Different from the depth-varying parallel data model, we extract all the time steps and plot them as vertical axes. Since in most cases, the number of time steps is much higher, and thus more axes would be generated. In this model, parallel axes represent time steps ranging from the beginning time to the ending time. However, not all the propagation records continually cover all the time steps. Most of the propagation paths just jump from one time step to another. Therefore, we need to deal with this "jumping" and the saltation elegantly. For those axes on which there are no propagation occurrences, we simply ignore these axes and connect those neighboring propagation instances together. The visualization plot of the same dataset is shown in Fig. 5.

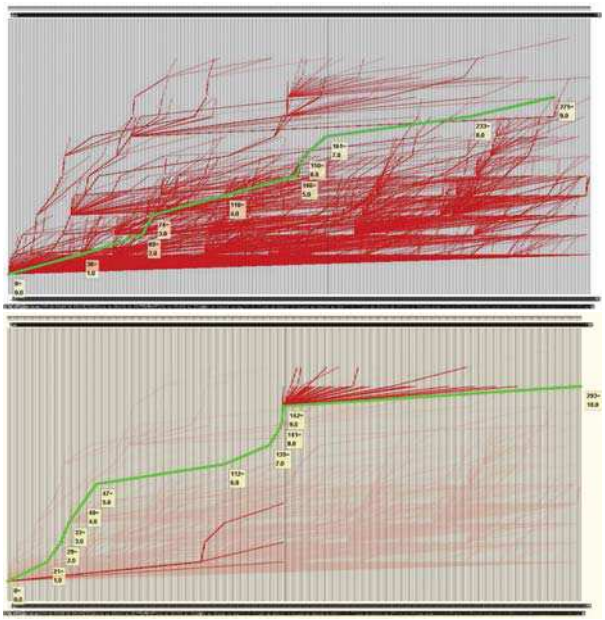

Figure 5: Time-varying parallel data visualization (After dealing with "jumps") and interactive exploration

In Fig. 5, it is easy to detect how the depths of propagation records distribute on a single axis from bottom to up, and from left to right, and thus we can get an overview of how all the propagation records change their paths as the time goes by. Meanwhile, we can add interactions to this plot and facilitate analysts to conveniently understand the retweeting propagation dataset better.

5.3 Hough-scatter Diagram

In this section we consider the speed of propagation and the patterns in the retweeting propagation of the depth-varying parallel data visualization plot. The traditional scatter plot only serves to represent dataset according to the distribution of values on dimensions, rather than present the underlying patterns in the dataset. By Hough transformation [20], the trend of lines between two neighboring axes can be easily detected. In our case, the transformation from the depth-varying parallel data model to the Hough plot makes the propagation speed apparent. In addition, in order to detect the "propagation clusters" in the Hough transformation plot, we add a density-based pixel plot, indicating the propagation summits from one depth to the successive one. Fig. 6 demonstrates the density-based pixel plot.

In Fig. 6, the upper part is the traditional scatter plot with density and the lower part is the Hough transformation plot with density. The green line indicates the "missing" values and we fill up them with -1. By observing the neighboring space between two adjacent axes, the "diagonal phenomenon" is common for scatter plot and the "horizon phenomenon" happens frequently. The reason for explaining this phenomenon is that a large number of the retweeting propagations happen instantly. In the upper panels, vertical clusters of red points can be witnessed as they are propagation summits.

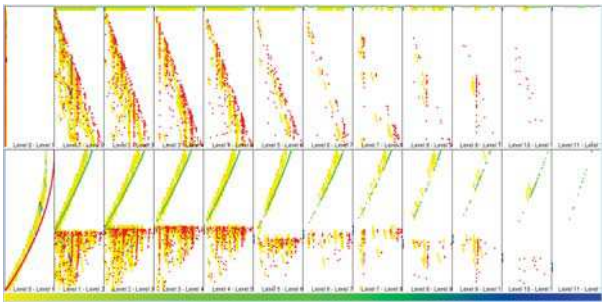

Figure 6: Density-based pixel plot for scatter plot and Hough plot. The green line indicates the "missing" values and the red points are the places that propagations occur

6. INTERACTIVE FEATURES

We combine several necessary interactive features to help users carry out those tasks and understand the properties of retweeting propagation networks.

Depth Filtering: Due to the large scale of certain retweeting propagation networks, we add a depth filtering strategy to control the data loading process and make the system perform smoothly and easily (Fig. 7).

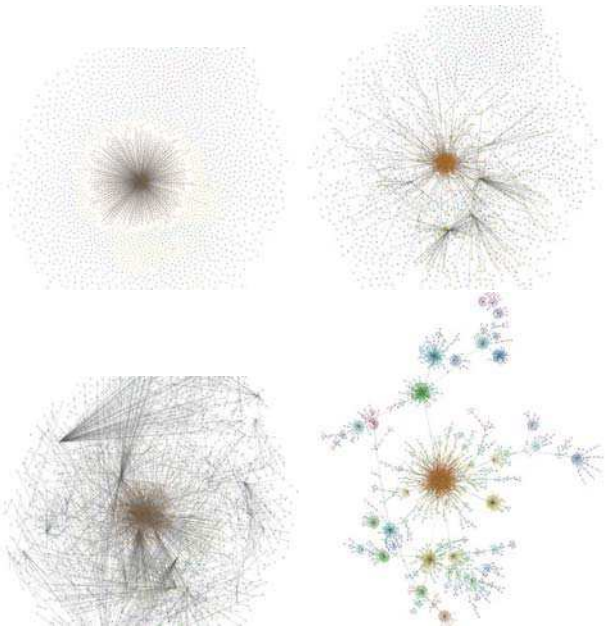

Figure 7: Depth controlling incremental visualization ($depthValue$ = 1, 2, 5, 12)(Only display edges of nodes of which the depth is 1, 2, 5, and 12)

Retweeting Filtering/Time Controlling: We provide users with freedom to adjust the retweeting number slider to filter out those nodes whose retweeting number is less than the threshold (see Fig. 8). As the network evolves with time, it is also quite intuitive to add a time slider to control the evolution of retweeting propagation graph, as shown in Fig. 9.

Lens: The lens lifts the focused items up while keeping other items staying still in the view plane. Meanwhile,

Figure 8: Retweeting filtering visualization (From left to right: $retweetingNumber \geq 0$, $retweetingNumber \geq 1$ and $retweetingNumber \geq 5$ at the final time step)

Figure 9: Time controlling visualization (From left to right: $time$ = 60, 150, and 293 minutes after the beginning of propagation)

users have the freedom to adjust the definition of focused items, such as changing the threshold retweeting number. Fig. 10 demonstrates the usage of lens to illustrate the focused retweeting propagation instances.

Propagation/Depth Clustering: We visualize this group information by restraining the boundary of all the nodes in the group and assign different colors to each group. The same goes to depth clustering. Fig. 11 is the visualization results by clustering retweeting propagation groups and depths. By turning on the group cluster displaying, we can clearly the boundary of each group and the information of group size. The depth displaying clearly shows the distribution of nodes within each depth.

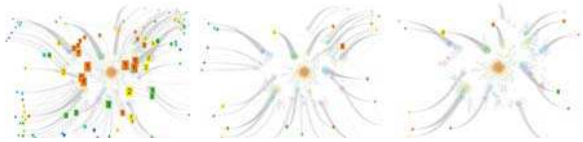

Figure 10: Lens: Lift the focused items up and keep others in the original view plane. (From left to right: threshold retweeting number is 5, 15, and 50, respectively)

7. CASE STUDY

7.1 Viral Online Marketing

In the first case, we collect the retweeting propagation records of a message published by a famous hotpot restaurant. This retweeting propagation dataset covers a large number of propagation records and lasts for a fairly long time, about ten days. But the number of its followers in the follower-fan network is limited. The success of micro-marketing in this case has attracted the attention of the relevant departments. We first plot the statistics of this

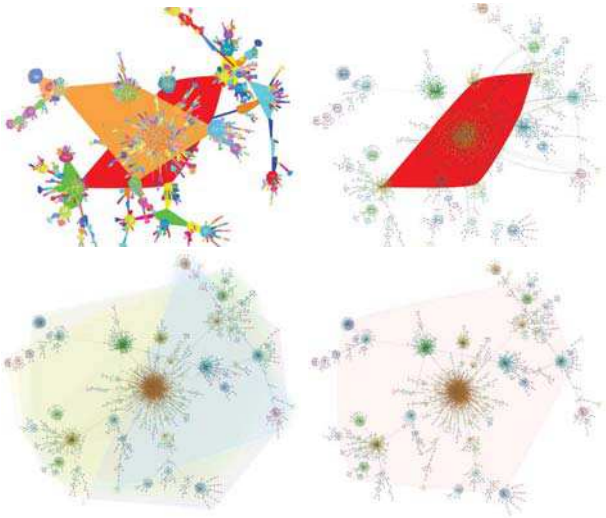

Figure 11: Visualization results by clustering retweeting groups and show different propagation coverage with different depths

retweeting propagation dataset by investigating in four aspects: occurrence along with time, occurrence along with depth, occurrence along with the number of retweeting, and occurrence along with the group size, respectively, as shown in Fig. 12. Although the retweeting propagation lasts for a long time, and extensive explosions are witnessed in several summits, such as after 250 minutes, 750 minutes, etc. after the beginning time. A large portion of propagation records are distributed in the first level, resulting in a large size of the first group.

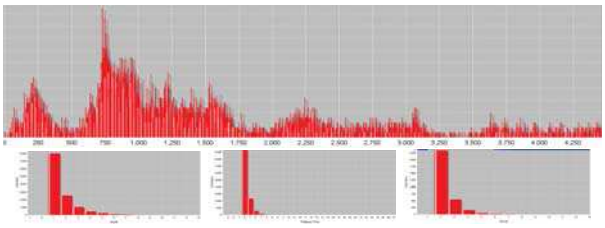

Figure 12: Statistics of the first retweeting propagation dataset

Next, we employ the Hierarchy Aware Force-Directed layout to plot all the records. We slightly adjust the time slider to watch the animation from the beginning to the current time and get the following four layouts after different lasting time (Fig. 13). In the final visualization, we then increase the value of depth slider to add edge restraint. The fifth and sixth plots of Fig. 13 show the initial step and the final result by adding edge restraint. Depth information of each vertex is encoded with the color of each vertex. Therefore, vertices with the same color belong to the same level. We turn on the display of group information and level displaying to see the boundary outline of the retweeting propagation structure, as shown in the seventh and eighth plots of Fig. 13. Most propagation paths are slim since most propagation instances just directly retweeted the original message from the root user.

By "Link + View" technique, we transform the fetched

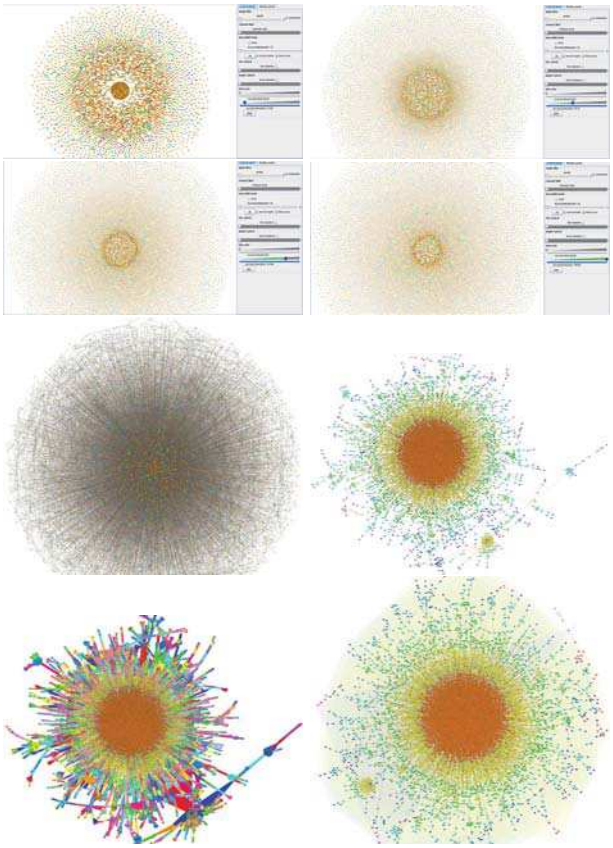

Figure 13: Hierarchy aware node-link diagram and some interactive techniques

information from the node-link diagram to the depth-varying parallel data model and emphasize the highly focused items with blue lines (the second plot of Fig. 14). We extract these key instances and find out that there are few intra-connections between them, telling us that those key nodes may not be strongly linked and may be connected by those unimportant nodes.

In the depth-varying parallel data model, we detect some interesting phenomenon: in propagation section 1 ranging from time 0 to time 3230 minutes (Fig. 15), the longest path has taken place. The longer the lasting time lasts, the shorter the propagation path is. One propagation summit happens in about 200 minutes, and it brings in three major propagation paths from depth 1 to depth 2, even to depth 8. The reason to explain for the long lasting time of propagation is that users on depth 1 continually retweeted the message after several breaks.

To take a deeper investigation about the relevant web users and their retweets, we select several interesting patterns, as illustrated in Fig. 16. All the selected patterns are propagated to distant depth of the retweeting propagation network and cover a wide range of web users. We are curious about why the whole propagation covers so many users while the number of its followers in the follower-fan network is limited. After crawling the relevant content of their retweets, we find that the effects of Circle of Friends have dramatic consequences. The web users who had retweeted

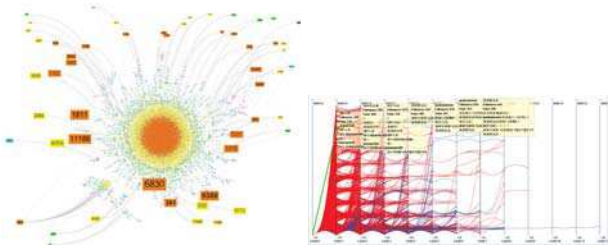

Figure 14: Link + View: Transform the fetched information from the node-link diagram to the depth-varying parallel data model and emphasize the highly focused items with blue lines

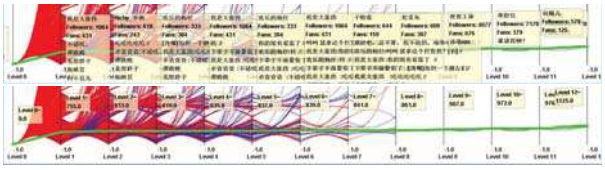

Figure 15: Propagation section 1 (Retweeting time displaying and related web users and their retweeting message)

this micro-blog were continually recommending this message to and interacting with their friends. By observing their mutual interactions, we conclude that it's quite important and necessary to promote discussion and interaction among web users.

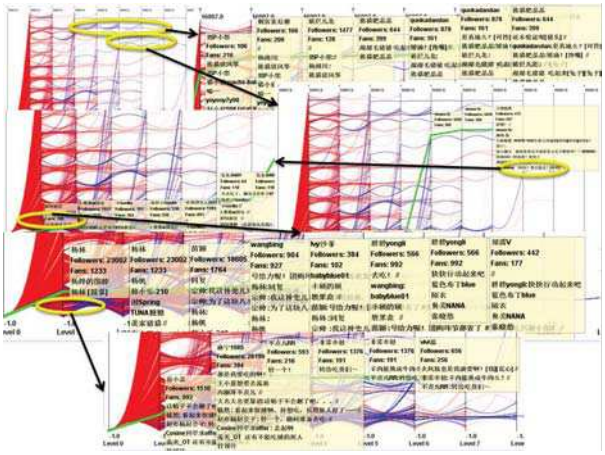

Figure 16: Relevant web users and their retweets

7.2 Event Detection

The second case is from a microblogging message published by a key figure. He published a message, recommending an article and a link pointing to this article. This dataset covers 1844 instances. The longest path is 6 and the average path is 3.09. It starts at 2011-11-20 07:37:02 and ends at 2011-11-20 23:58:43, which lasts for 981 minutes. We directly plot the node-link diagram, highlight those nodes whose retweeting number is greater than 15, and map them to a depth-varying parallel data plot, as shown in Fig. 17.

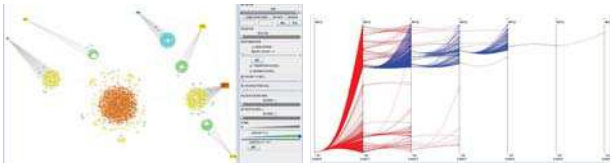

Figure 17: From left to right: node-link diagram and depth-varying parallel data visualization

In Fig. 17, the longest path is found in 647 minutes after the beginning of publishing the message. From depth 0 to depth 1, two distinguished parts are formed after a long time break. Therefore, we are interested in the following questions: (1) How to explain the two distinguished parts when propagation from depth 0 to depth 1? (2) Why does the longest path appear after a long time? In most cases, the first half part of the longest path always appears at the bottom of the visualization plot. (3) Why does not the propagation from depth 0 to depth 1 at an early time last to a deeper level?

To filter out those unimportant nodes and keep the main structure of the graph, we highlight the center of each cluster in a node-link diagram, map them to the depth-varying model and then employ our simplified layout algorithm to keep the top two accumulation degree of each depth. We fetch the related information of users and tweets (Fig. 18). Then, we find out the reasons: The original message is about article recommendation published by root user A. This article was written by user B. However, the URL link in the original message was incorrect until in about 647 minutes when the article's author—user B retweeted this message and corrected it with the right URL link, which led to a great rise of propagation from depth 0 to depth 1 again, since this corrected message appeared on the retweeting webpage of root user A. At the early time from depth 0 to depth 1 the propagation did not reach to a deeper level, although there was a key figure C having retweeted this message and brought in some propagations. However, we cannot find that that key figure's retweeting record by API connection. Therefore, we conclude that user C must have deleted this original message that contains the wrong URL link, resulting in no more propagations. And the propagations that passed the article author—user B increased significantly and formed the longest path in the overall retweeting propagation network.

Then, we are interested in how the propagation paths split, emerge and vanish with the spread of the scope of the constantly propagation and deepening. In the scatter plot and Hough plot, we can clearly see the changes of propagation paths. We select the interested pattern with a blue circle that presents a long propagation path, and this will simultaneously reflect in other relevant panels and depth-varying data view. The selected pattern happened from propagation depth 1 to depth 2 splits into three parts. Two of them are still spreading and the other part has vanished, as show in Fig. 19 (top). As we have demonstrated previously, the "diagonal phenomenon" in the scatter plot and "horizon phenomenon" in the Hough plot are apparent, which indicate the instantly retweeting behaviors from parent nodes to child nodes. We select the "horizon line" in the Hough plot and check out the corresponding patterns. In this case, two thirds of the propagation instances from depth 2 to depth 3 are due to the instantly retweeting behaviors from depth 1

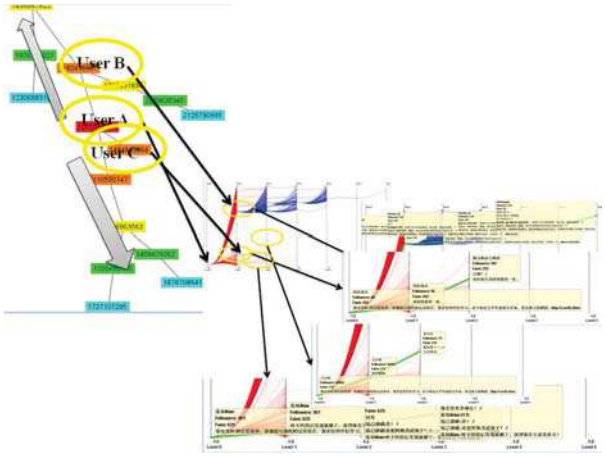

Figure 18: Simplified layout with displaying the identification of each vertex and the relevant patterns of them

to depth 2 and another one third of propagation is caused by a retweeting behavior delay (Fig. 19 (bottom)).

After a long time break, the blue line in "Level 1-Level 2" panel becomes the major propagation in the next panel and then breaks into three major branches (Fig. 20 (left)). In the time-varying parallel data visualization plot, the distribution of depth and time scale is apparent. The structure of this graph presents a major trend: In about 647 minutes after the beginning of propagation, retweeting from the root user still occupies a lot, much denser than propagations from depth 1 to depth 2, or depth 2 to higher depths (Fig. 20 (right)).

7.3 Interview with Data Analysts and Technical Support Engineer

We conducted extended one-on-one interviews with twenty data analysts and one technical support engineer with very strong domain expertise. Most of the data analysts were majoring in statistics. The technical support engineer is responsible for providing technical assistance between the service platform and our visualization system.

All the data analysts were intrigued by the interactive features that our system provides for examining propagation patterns and highlighting those key propagation instances. Before adopting our system, they used a simple tool which generates a static image by directly plotting all the propagation records. They complained that the previous tool can only deal with relatively small data and does not support any user interactions. For our system, the data analysts were attracted to the interactivity and felt that the multi-layouts provided significant value. They liked the mechanism of the system that the displaying of propagation nodes is triggered by interactions or time-steps, instead of displaying all of them. They thought that the improved Hierarchy Aware Force-Directed layout is more in line with their intuitions.

Referring to the detailed designs of our system, one data analyst said "it gives me an intuitive way to examine propagation instances just by taking a glance at the color encoding distribution". And the animation provided in the system makes the data analysts convenient to control their interested time-steps. Another data analyst responsible for

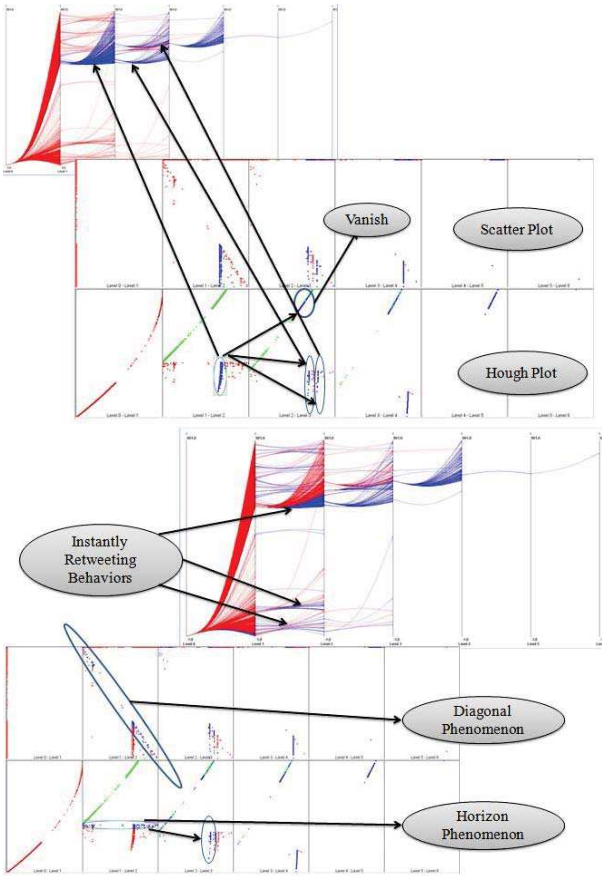

Figure 19: Propagation splitting and instantly retweeting behaviors presentation

analysis of micro-grouping product, was especially interested in the drag-and-drop nature of the technique which she felt provided a "very intuitive interface" for manipulating the positions of each micro-groups. She liked the grouping design encoded by different color clusters and the intra-and-inter connections for her to analyze and navigate.

What they liked most, is the two proposed models: the depth-varying and the time-varying parallel data model which intuitively present the dynamic attributes. "I can get a lot of information", said one of the data analysts, "these two models represent the distribution of time, depth, propagation intervals and even propagation clusters clearly". When asked whether this system helped them understand the underlying information of propagation data and improve their work efficiency, they responded "of course". We also design

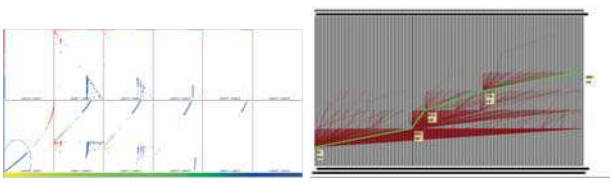

Figure 20: Hough scatter plot and time-varying parallel data visualization

several objective and quantifiable indicators. We grade each indicator by scoring from one to ten. Fig. 21 shows the comparison by averaging their feedbacks.

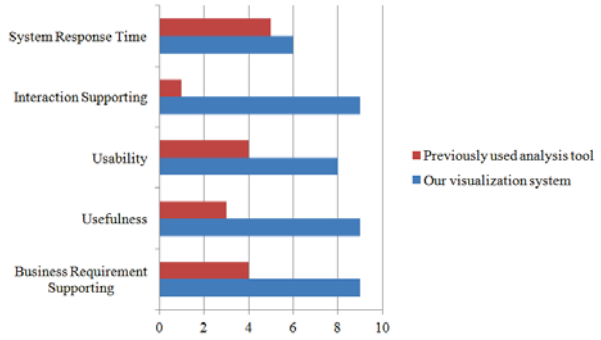

Figure 21: Quantifiable indicators to compare the performance between the previously used analysis tool and our visualization system

8. CONCLUSION AND FUTURE WORK

In this paper we have investigated how information propagation in a specific microblogging platform evolves from a visual perspective. Scenario-oriented layouts have been proposed to map the multi-attribute propagation network into appropriate visual elements. We have also introduced two parallel data models to depict the dynamic retweeting propagation graph in a single image, which has clear advantages over the traditional presentation for dynamic retweeting propagation graph. The proposed methods and several interactive techniques have been integrated into our visual mining system, which has been evaluated and used in the Weibo department of Sina Corporation. In the future, we will exploit some NLP techniques, such as opinion mining and clustering to analyze the content of messages and provide a deeper understanding of propagation.

9. ACKNOWLEDGMENTS

This work was supported by Tsinghua-Sina Project, National Science Foundation of China(61272225,51261120376) and Chinese 863 Program(2012AA041606,2012AA040902).

10. REFERENCES

[1] J. Barnes and P. Hut. A hierarchical 0 (n log iv) force-calculation algorithm. *nature*, 324(4), 1986.

[2] M. Burch, C. Vehlow, F. Beck, S. Diehl, and D. Weiskopf. Parallel edge splatting for scalable dynamic graph visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2344–2353, 2011.

[3] N. Diakopoulos and D. Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1195–1198. ACM, 2010.

[4] S. Diehl and C. Görg. Graphs, they are changing. In *Graph Drawing*, pages 23–31. Springer, 2002.

[5] P. Eades. A heuristic for graph drawing. *Congressus numerantium*, 42:149–160, 1984.

[6] Y. Frishman and A. Tal. Online dynamic graph drawing. *Visualization and Computer Graphics, IEEE Transactions on*, 14(4):727–740, 2008.

[7] T. Fruchterman and E. Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164, 1991.

[8] D. Gaffney. # iranelection: quantifying online activism. 2010.

[9] C. Görg, P. Birke, M. Pohl, and S. Diehl. Dynamic graph drawing of sequences of orthogonal and hierarchical graphs. In *Graph Drawing*, pages 228–238. Springer, 2005.

[10] P. Grabowicz, J. Ramasco, E. Moro, J. Pujol, and V. Eguiluz. Social features of online networks: the strength of weak ties in online social media. *Arxiv preprint arXiv:1107.4009*, 2011.

[11] C. Ho. Modeling and visualizing information propagation in a micro-blogging platform (pdf). 2011.

[12] A. Inselberg and B. Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proceedings of the 1st conference on Visualization'90*, pages 361–378. IEEE Computer Society Press, 1990.

[13] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009.

[14] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Information processing letters*, 31(1):7–15, 1989.

[15] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[16] J. Leskovec. Social media analytics: tracking, modeling and predicting the flow of information through networks. In *Proceedings of the 20th international conference companion on World wide web*, pages 277–278. ACM, 2011.

[17] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[18] D. Shamma, L. Kennedy, and E. Churchill. Conversational shadows: Describing live media events using short messages. *Proceedings of ICWSM*, 2010.

[19] E. Sun, I. Rosenn, C. Marlow, and T. Lento. Gesundheit! modeling contagion through facebook news feed. *Proc. ICWSM*, 9, 2009.

[20] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnork, and D. Keim. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 59–66. IEEE, 2009.

[21] T. Von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. Van Wijk, J. Fekete, and D. Fellner. Visual analysis of large graphs. *Proceedings of Euro-Graphics: State of the Art Report*, 2, 2010.