SocialWave: Visual Analysis of Spatio-temporal Diffusion of Information on Social Media

GUODAO SUN, Zhejiang University of Technology
TAN TANG, Zhejiang University
TAI-QUAN PENG, Michigan State University
RONGHUA LIANG, Zhejiang University of Technology
YINGCAI WU, Zhejiang University

Rapid advancement of social media tremendously facilitates and accelerates the information diffusion among users around the world. How and to what extent will the information on social media achieve widespread diffusion across the world? How can we quantify the interaction between users from different geolocations in the diffusion process? How will the spatial patterns of information diffusion change over time? To address these questions, a dynamic social gravity model (SGM) is proposed to quantify the dynamic spatial interaction behavior among social media users in information diffusion. The dynamic SGM includes three factors that are theoretically significant to the spatial diffusion of information: geographic distance, cultural proximity, and linguistic similarity. Temporal dimension is also taken into account to help detect recency effect, and ground-truth data is integrated into the model to help measure the diffusion power. Furthermore, SocialWave, a visual analytic system, is developed to support both spatial and temporal investigative tasks. SocialWave provides a temporal visualization that allows users to quickly identify the overall temporal diffusion patterns, which reflect the spatial characteristics of the diffusion network. When a meaningful temporal pattern is identified, SocialWave utilizes a new occlusion-free spatial visualization, which integrates a node-link diagram into a circular cartogram for further analysis. Moreover, we propose a set of rich user interactions that enable in-depth, multi-faceted analysis of the diffusion on social media. The effectiveness and efficiency of the mathematical model and visualization system are evaluated with two datasets on social media, namely, Ebola Epidemics and Ferguson Unrest.

CCS Concepts: • **Human-centered computing → Visual analytics**;

Additional Key Words and Phrases: Spatio-temporal visualization, information diffusion, social media visualization

The work is supported by National 973 Program of China (2015CB352503), National Natural Science Foundation of China (61602409, U1609217, 61502416), the Fundamental Research Funds for Central Universities (2016QNA5014), the research fund of the Ministry of Education of China (188170-170160502), 100 Talents Program of Zhejiang University, Zhejiang Provincial NSFC (No. LR14F020002), and the Open Projects Program of Key Laboratory of Ministry of Public Security based on Zhejiang Police College (2016DSJSYS003).

Authors' addresses: G. Sun and R. Liang, College of Information Engieering, Zhejiang University of Technology, No. 288 Liuhe Road, Xihu District, Hangzhou, China; emails: {guodao, rhliang}@zjut.edu.cn; T. Tang and Y. Wu (corresponding author), State Key Lab of CAD & CG, College of Computer Science, Zhejiang University, No. 866 Yuhangtang Road, Xihu District, Hangzhou, China; emails: tangtan1012@gmail.com, ycwu@zju.edu.cn; T. Q. Peng, Department of Communication, Michigan State University, 404 Wilson Road, East Lansing, MI, USA; email: winsonpeng@gmail.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2017 ACM 2157-6904/2017/10-ART15 $15.00
https://doi.org/10.1145/3106775

ACM Reference format:
Guodao Sun, Tan Tang, Tai-Quan Peng, Ronghua Liang, and Yingcai Wu. 2017. SocialWave: Visual Analysis of Spatio-temporal Diffusion of Information on Social Media. *ACM Trans. Intell. Syst. Technol.* 9, 2, Article 15 (October 2017), 23 pages.
https://doi.org/10.1145/3106775

1 INTRODUCTION

With increasing penetration of social media and rapid prevalence of mobile technologies, popular topics on social media can be diffused among users from different geolocations in either synchronous or asynchronous manners (Kamath and Caverlee 2013; Leskovec et al. 2009). It has been a popular practice to pool together users from different geolocations in existing studies on information diffusion (Zheng et al. 2014). However, this practice does not take into account the differences on behavioral patterns, linguistic traditions, and cultural values between users in different societies (Dodds et al. 2015; Ronen et al. 2014). These issues limit our theoretical understanding of the subtle process underlying information diffusion on social media. Moreover, an increasing demand for understanding the spatio-temporal characteristics of information diffusion on social media is observed in various practical scenarios (Bosch et al. 2013; Kamath et al. 2013). For instance, to efficiently contain the diffusion of rumors on social media, government policy-makers need to identify the critical geographic areas where and key time windows when the rumors originate from, become viral, and fade out, which can further help them disseminate the truth to users in those critical areas and at optimal time points (Cao et al. 2012; Sakaki et al. 2010).

Although theoretically significant and practically necessary, measuring and understanding the spatio-temporal characteristics of information diffusion on social media is a daunting challenge in empirical research. Various sets of factors, including temporal, spatial, and user characteristics, are found to affect information diffusion on social media (Leskovec et al. 2009; Liben-Nowell et al. 2005; McPherson et al. 2001). Thus, the first challenge is how to integrate these factors in a logical and seamless way into our measurement of the spatio-temporal characteristics of information diffusion. Moreover, as the spatio-temporal characteristics of information diffusion may exhibit different patterns when different temporal and spatial granularities are adopted, how to handle the multi-granularity characteristics in space and time is the third challenge. Last but not least, how to visualize the spatio-temporal characteristics of information diffusion in an intuitive, interactive, and insightful way is the fourth challenge we have to face.

In recent years, empirical studies have been conducted to examine the spatio-temporal characteristics of information diffusion on social media (Bosch et al. 2013; Cao et al. 2012; Kamath and Caverlee 2013; Kamath et al. 2013). These studies focused either on mapping the spatio-temporal diffusion of raw information on social media (e.g., microblogging messages) (Bosch et al. 2013; Cao et al. 2012), or on predicting the spatio-temporal diffusion of information with probabilistic models (Kamath and Caverlee 2013; Kamath et al. 2013). However, there are several limitations with existing studies. First, these studies implicitly assume that users from different geographic areas will be influenced by one another in the process of information diffusion. They did not take the essential temporal, spatial, and user characteristics into account to explicitly quantify the dynamic influence between users in different geographic areas. Second, most of the existing studies examined information diffusion in single spatial granularities (e.g., countries, states, and cities) or temporal granularities (e.g., week, day, and hour), which leads to different patterns and causes the difficulty, if not the impossibility, to compare research findings across studies. Third, most of the existing studies on visualization of information diffusion do not allow for large-scale visual spatio-temporal exploration (Sun et al. 2014; Viégas et al. 2013; Wu et al. 2014).

In this study, a model called dynamic social gravity model (SGM) is developed to quantify the spatio-temporal dynamics of information diffusion on social media. The dynamic SGM is built by extending the classic gravity model, which is widely adopted in linguistics, economics, and demography research (Anderson and van Wincoop 2003; Sgrignoli et al. 2015; Trudgill 1974). Compared with the classical gravity model, the dynamic SGM has three advantages. First, it simultaneously takes into account the three important dimensions in spatial diffusion: namely, *geographic distance* (Liben-Nowell et al. 2005; Scellato et al. 2010), *cultural proximity* (Hoftede et al. 2010), and *linguistic similarity* (McPherson et al. 2001; Trudgill 1974), which allow researchers to directly capture the spatial interaction between social media users in information diffusion. Second, the dynamic SGM explicitly takes the temporal dimension into account by allowing some factors (i.e., population size and linguistic similarity) to vary over time, which can help us detect the *recency effect* (Leskovec et al. 2009; Macskassy and Michelson 2011) in information diffusion and advance our understanding of the complex dynamics underlying information diffusion in a more comprehensive way. Moreover, the spatial dimension in the dynamic SGM can be scaled to different granularities to support multi-granularity analysis. Third, ground-truth dataset can be integrated into the dynamic SGM to help precisely measure the diffusion power using regression analysis (Liu et al. 2014; Viboud et al. 2006).

With the dynamic SGM, we further develop a visual analytical system called SocialWave to facilitate the exploration and analysis of the spatio-temporal diffusion of information on social media. A timeline visualization, which considers the spatial characteristics of the diffusion network, is provided to allow users to quickly identify critical time periods when the diffusion is significant. We propose an innovative visualization that integrates a node-link diagram into a multi-scale circular cartogram (Dorling 1996) to create a flexible, expressive, and occlusion-free visualization. This innovation aims to represent the complex dynamics of the spatio-temporal diffusion captured by SGM for any critical time periods. The simplicity and distinction of the circular cartogram, and the intuitiveness of the node-link diagrams are delivered with our design. Visual clutter is further reduced through iterative exertion of repulsive and attractive forces on the nodes and edges. The key contributions of this work are as follows:

—We propose an extended model that quantitatively characterizes the spatio-temporal diffusion of information on social media under multiple granularities.
—We design and develop SocialWave, a visual analytics system for interactive visual exploration and summarization of the complex spatio-temporal diffusion of information on social media.
—We provide empirical findings based on two large-scale social media datasets to test the effectiveness of our theoretical model and the visual analytics system.

2 RELATED WORK

This section reviews related work on information diffusion model, visualization of the diffusion process, and graph visualization.

Modeling of Information Diffusion. Extensive efforts have been made to model information diffusion on social media: from traditional linear threshold and independent cascade model (Easley and Kleinberg 2010) to new advanced models involving different diffusion mechanisms such as user's influence ability (Ho et al. 2011), adoption probability (Romero et al. 2011), or to predict the propagation (Cheng et al. 2014). More recently, the development of location-based services provided by social media sites has motivated research on spatio-temporal analysis of information diffusion on social media (Caverlee et al. 2013). The following factors are regarded as crucial in modeling spatio-temporal diffusion of information: geographic distance (Liben-Nowell et al. 2005;

Scellato et al. 2010), recency effect (Leskovec et al. 2009; Macskassy and Michelson 2011), and cultural proximity and linguistic similarity (Hoftede et al. 2010; McPherson et al. 2001; Trudgill 1974). These factors, however, are not fully and adequately considered in recent research. Leskovec et al. (2009) and Macskassy and Michelson (2011) merely focused on the recency effect to model information diffusion. Kamath and Caverlee (2013) proposed a probabilistic model to measure and predict the spatio-temporal distribution and popularity of hashtags. This work only integrates spatial influence (geographic distance) and community influence (linguistic similarity) into the modeling of diffusion.

These studies could not quantify the diffusion power among different places and are unable to explore the multi-level spatio-temporal diffusion with different spatial and temporal granularities. In our study, we propose to extend the classic gravity model (see Section 4.1) to characterize the spatio-temporal diffusion of information. The classic gravity model is commonly used in communication, economic, and demography field to explain and predict the extent of flow among places (Anderson and van Wincoop 2003; Sgrignoli et al. 2015; Trudgill 1974). Our model is scalable and supports multi-level analysis. Compared with existing models, our SGM can fully consider the effect of geographic distance, recency effect, and cultural proximity and linguistic similarity.

Visualization of Information Flow. Various visualization methods have been proposed to present and analyze the spatial and/or temporal information diffusion on social media, while most of the work focuses on the temporal dimension of information diffusion (Sun et al. 2014; Viégas et al. 2013; Wu et al. 2014; Times 2011; Guardian 2011; Stefaner 2013; Zhao et al. 2014). In the present work for visualizing spatio-temporal information diffusion on social media, Cao et al. (2012) employed a sunflower metaphor to visualize the spatio-temporal propagation of raw tweets and sentiment. Ho et al. (2011) visualized the geographical distribution of the tweets and sentiment related to a specific event on a map. Marcus et al. (2011) directly marked the geographic propagation on maps to visualize the geographic influence. However, most of these works use geographical maps as the background and overlay diffusion process directly on the maps, which could introduce serious visual clutter and occlusion, particularly for large diffusion graphs. This issue can considerably hinder the exploration of spatio-temporal diffusion. In contrast, our work, SocialWave, leverages the advantages of the simplicity and distinction of circular cartogram (Dorling 1996) and the intuitiveness of node-link diagrams. Visual clutter and occlusion issues can be properly addressed through iterative exertion of repulsive and attractive forces on the nodes and edges.

Graph Visualization. Graph visualization has attracted considerable attention in recent years. We focus our discussion on visual clutter reduction of node-link diagrams that are closely related to our work. Interested readers can refer to excellent surveys and books for a complete review of graph visualization techniques (Herman et al. 2000; Kaufmann and Wagner 2001).

Researchers have proposed different strategies such as edge displacement, node clustering, and graph sampling to reduce visual clutter of graph visualization. Graph sampling methods (Leskovec and Faloutsos 2006; Rafiei 2005), such as node sampling and edge sampling, have been widely used to obtain a sampled representative graph for exploration. However, these methods could disregard important diffusion patterns that are not sampled. Node clustering (Eades et al. 1997; Kaufmann and Wagner 2001) has also been used to simplify a large, complex graph by clustering similar nodes. However, solely relying on node clustering is not sufficient, because the remaining nodes and edges with varying sizes and widths may still occlude each other.

Researchers have introduced many edge displacement methods such as hierarchical edge bundling (Holten 2006), geometry-based edge bundling (Cui et al. 2008), force-directed edge bundling (Holten and Van Wijk 2009), divided edge bundling (Selassie et al. 2011), and flow map (Phan et al. 2005; Verbeek et al. 2011) to group edges and reduce visual clutter. The existing edge bundling methods have been successfully applied to reduce visual clutter of general graphs.

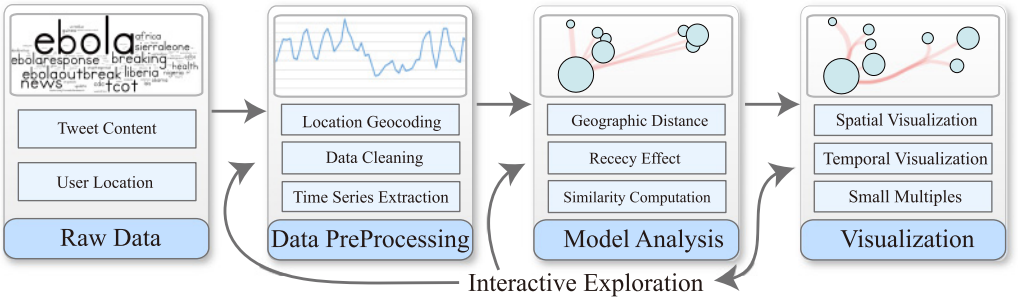

Fig. 1. Overview of the system SocialWave, which includes three major parts: data preprocessing, model analysis, and interactive visualization.

However, they cannot be directly applied to visualize the complex graph (captured by our proposed spatio-temporal diffusion model), namely, a dynamic, directed, dyadic full graph with nodes and edges that have different sizes and widths. Because these methods ignore either the edge width or the node size, which could easily lead to occlusion and clutter issues. Ambiguity free edge bundling (Luo et al. 2012) has been proposed to avoid edge ambiguity by routing edges through nearby empty cells. This method works for locally dense graphs but may fail to reduce clutter when nodes are overlapped or no empty cell exists near ambiguous edges.

Our work employs a novel method that takes into account the node size and edge width to rearrange the graph layout and improve its readability. The main idea is similar to the circular cartogram that relaxes the geolocation constraint of nodes while still maintaining users' geographical mental map. More specifically, our method allows the nodes to be slightly shifted around their original places while preserving the relative spatial relationships with their neighbors. In contrast to the circular cartogram, our method considers both the weighted nodes and the weighted edges to exert the repulsive and attractive forces iteratively, such that adequate space could be reserved to bundle edges and avoid the overlaps while preserving users' mental map of the geographical reference of the nodes in the original map.

3 SYSTEM OVERVIEW

SocialWave includes three major parts, namely, data preprocessing, model analysis, and interactive visualization. Figure 1 illustrates the system overview. The data preprocessing part cleans Twitter data and uses Google Geocoding API[1] to geocodes the user location for each tweet by approximating the latitude and longitude of the corresponding user from the place provided in his/her profile. This strategy is commonly recognized in performing geography-related exploration tasks (Sakaki et al. 2010). A high-performance information retrieval engine, Apache Lucene,[2] is adopted subsequently to facilitate text indexing and searching. Time series for hashtag occurrence in different locations at different times are extracted through the engine. The model analysis part is fed with various time-varying and time-invariant data such as the temporal occurrence of hashtags for each place, temporal linguistic similarity, and cultural proximity among different places. Our SGM can quantitatively measure the dynamic diffusion power among different places. The visualization part is further fed with the output of the model analysis part.

[1] https://developers.google.com/maps/documentation/geocoding/.
[2] https://lucene.apache.org/.

It visualizes the dynamic information diffusion in both temporal and spatial dimension with coordinated views to enable interactive investigative analysis.

4 MODEL

This section introduces the classic gravity model and then describes our dynamic SGM to quantify the spatio-temporal properties of information diffusion on social media.

4.1 Classic Gravity Model

Gravity model has been hailed as one of the most popular paradigms in studying spatial interaction behavior of human populations in various research areas (Anderson and van Wincoop 2003; Trudgill 1974; Sgrignoli et al. 2015). The classical gravity model posits that the interaction between two populations is directly related to their population size and inversely related to the extent of separation between them (Isard and Bramhall 1960). There are two major challenges in applying the classic gravity model to examine the spatio-temporal properties of information diffusion.

The first challenge is how to operationalize the extent of separation between populations. The most intuitive operationalization of the extent of separation is the geographic distance between two populations. Although the cyberspace has been touted to be boundless, human interaction behavior, such as befriending and communication, is recognized as geographically determined (Liben-Nowell et al. 2005; Peng et al. 2015). In addition, other dimensions relevant to specific research contexts should be considered to achieve a valid and reliable understanding of spatial interaction behavior (Sen and Smith 1995). In this study, the diffusion of information among Twitter users from different geolocations should matter with the separation in three aspects, namely, geography, culture, and linguistic pattern. The second challenge is the static nature of classic gravity model, which assumes that all the parameters included are time-invariant. Most of the empirical studies on classic gravity model focused on providing a snapshot description of interaction behavior in specific time periods (Sen and Smith 1995), although interaction behavior between human populations is an ongoing process. These static descriptions assumed that all the parameters included in the model were time-invariant. In the current study, we explicitly integrate the temporal variations of relevant parameters into the gravity model and propose an extended model, which can allow us to uncover the dynamics of interaction behavior involving significant structural changes.

4.2 Dynamic Social Gravity Model

With the aforementioned challenges, we propose an extended model, called dynamic social gravity model, that quantifies the diffusion of information from users in location i to users in location j at time t with a combination of following factors, namely, *salience of information* among users in locations i and j, *geographic distance* between locations i and j, *recency effect* for information adoption between location i and j at two adjacent time t and t-1, and *cultural proximity* and *linguistic similarity* between users in locations i and j. These factors are considered practically necessary and theoretically significant in characterizing the diffusion pattern of information across different geolocations over time. In this study, we use hashtag as the elementary propagation unit in modeling spatio-temporal diffusion of information on social media. With the consideration of above crucial factors, we model the diffusion power from users in location i to users in location j at time t with respect to a hashtag h as follows:

$$R^{t}_{i \xrightarrow{h} j} = S^{\gamma}_{i(t-1),j(t)} \frac{P^{\tau_1}_{i(t-1)} P^{\tau_2}_{j(t)}}{d^{\rho}_{i,j}}, \tag{1}$$

where

$$S_{i(t-1),j(t)} = \frac{1}{c_{i,j}} \frac{\sum_w min(P_{w,i(t-1)}, P_{w,j(t)})}{\sum_w max(P_{w,i(t-1)}, P_{w,j(t)})}. \quad (2)$$

In Equation (1), $P_{i(t-1)}$ and $P_{j(t)}$ are the salience of hashtag h observed among users in location i and j at time *t-1* and *t*. The exponential parameters τ_1, τ_2 , ρ, and γ moderate the dependence of diffusion strength on the salience of hashtag h observed in location i and j, the geographical distance and semantic similarity between the two locations. The salience of a hashtag in a location is computed by dividing the number of occurrences of the hashtag in that location by the total number of occurrences of the hashtag across all the locations. $P_{i(t-1)}$ and $P_{j(t)}$ could be scaled to multiple hashtags by adding the number of occurrences of the hashtags observed in corresponding locations respectively. Other terms in our dynamic SGM are elaborated as follows with respect to the aforementioned factors:

Geographic Distance. Information and communication technologies have been touted to break the geographic boundaries in human society (Cairncross 2001). However, recent empirical findings imply that the geographic distance still plays a significant role in friendship formation and information diffusion (Liben-Nowell et al. 2005; Peng et al. 2015; Scellato et al. 2010). The greater the geographic distance between two geolocations, the less interaction users from both locations tend to have. Thus, a distance decay function (i.e., $d_{i,j}^{\rho}$) is an indispensable component in our model to estimate the diffusion power between the users from different locations. $d_{i,j}$ in Equation (1) represents the geographic distance between location i and j. The exponent value ρ are supposed to be positive. The geographic distance between two geolocations is calculated using the Haversine formula, a widely used method in calculating great-circle distance on Earth's sphere at two points with specific longitude and latitude.

Recency Effect. A temporal order is implied in all diffusion processes where information (i.e., hashtags in our study) spreads from users in location i to users in location j. Therefore, it is theoretically appealing and methodologically necessary to take the temporal dimension into account when examining the influence of users from geolocation i on users from geolocation j. *Recency effect* refers to a cognitive bias that people are more likely to recall the information they have been recently exposed to. *Recency effect* has been found to be prominent in information diffusion mostly due to the overloaded information environment on social media (Leskovec et al. 2009; Macskassy and Michelson 2011). Therefore, it is reasonable for us to assume that users in location j at time t will be influenced by those in location i at time *t-1*. The "recency" is defined as a 4h time window in the study, since a 4h time window can provide adequate observations to capture users' temporal activities (Juster and Stafford 1991; Sun et al. 2014).

Cultural Proximity and Linguistic Similarity. With everything else being equal, information is more likely to spread between two populations when they have similar cultural background (McPherson et al. 2001; Trudgill 1974). Furthermore, users tend to adopt the information that is linguistically similar to what they used or were exposed before (Sun et al. 2014). In our model, we include a culturally weighted linguistic similarity (i.e., $S_{i(t-1),j(t)}$), which combines both cultural proximity and linguistic similarity in assessing the influence between users from different locations. We adopt a widely used framework called Hofstede's cultural dimensions theory (Hoftede et al. 2010) to measure the cultural proximity (i.e., $c_{i,j}$) between two geolocations. Six dimensions, namely, *power distance*, *individualism*, *uncertainty avoidance*, *masculinity*, *long-term orientation*, and *indulgence* are evaluated for each country based on Hofstede's model. The cultural proximity between two geolocations is calculated by computing the Euclidean distance of corresponding two feature vectors with respect to the six dimensions. In Hofstede's model, some countries are missing in the above six dimensions. We use the average value of the corresponding

Table 1. Parameters Estimation of #Ebola in a Global-scale Model

Parameter	Estimate	Standard Error	t Value	P Value
τ_1	0.14840***	0.01824	8.135	6.55e-16
τ_2	0.14075***	0.01849	7.611	3.90e-14
ρ	0.42729***	0.03744	−11.413	<2e-16
γ	0.22175**	0.05781	3.836	0.000128

Significance of parameter: *<0.01, **<0.001, ***<0.0001.

dimensions of neighboring countries to impute the missing values in Hofstedes cultural dimensions. For the state/province or an even finer level within a country, we assume that the culture is homogeneous within a country (i.e., $c_{i,j} = 1$). We further measure the linguistic similarity between users in two locations using the weighted Jaccard coefficient (Ioffe 2010). The weighted Jaccard coefficient (i.e., $\frac{\sum_w min(P_{w,i(t-1)}, P_{w,j(t)})}{\sum_w max(P_{w,i(t-1)}, P_{w,j(t)})}$) considers the sets of unique hashtags mentioned in the two locations $P_{w,i(t-1)}$ and $P_{w,j(t)}$, which represent the salience of hashtag w observed in geolocation i and j at time *t-1* and t, respectively. Geolocations that possess common hashtags with equal number of occurrences will have a linguistic similarity score of 1, and those that do not share any hashtag will have a score of 0.

4.3 Measuring Diffusion Power

The diffusion power between two locations takes the form of Equation (1), however, the parameters τ_1, τ_2, ρ, and γ are still unknown. Therefore, we propose a regression model by replacing the left side of Equation (1) with comparable dataset, which include the communication behaviour between different locations. In this study, we adopt a flight network dataset[3] as a ground-truth dataset to evaluate the proposed dynamic SGM. Flight network is commonly used to explicitly illustrate and acts as a regression analysis simulation dataset for the spatial communication among different locations with the form of gravity model (Liu et al. 2014; Viboud et al. 2006). Different spatial scale and selection of hashtag(s) may result in different parameters. For example, we can adopt the country level of flight network dataset to estimate the parameters for global events and state level to estimate the ones for domestic event. In Table 1, we illustrate the estimated parameters for the hashtag #Ebloa within a global-scale model as an example. t-test and P-value shows that all factors in the dynamic SGM are significantly associated with the real-world communication behaviours.

Moreover, we are interested in examining the overall interaction between users in two geolocations in the diffusion process regardless of the diffusion direction. Many extended gravity models for measuring spatial interaction assume that the influence between location i and j could be further decomposed into the addition of the influence from i to j and that from j to i (Trudgill 1974). Thus, following this strategy, we further model the overall diffusion power between location i and j at time t with respect to a hashtag h as follows:

$$R^t_{i \overset{h}{\leftrightarrow} j} = R^t_{i \overset{h}{\longrightarrow} j} + R^t_{j \overset{h}{\longrightarrow} i}, \tag{3}$$

where $R^t_{j \overset{h}{\longrightarrow} i}$ represents the diffusion power from location j to i at time t with respect to hashtag h and could be simply derived in the same way as expressed in Equation (1). The diffusion power

[3] http://www.openflights.org/.

between two locations can indicate the overall magnitude of the diffusion power between the locations and is useful in presenting the overall diffusion pattern among different geolocations at a given time.

5 VISUAL DESIGN

This section introduces the user requirements and research questions collected from domain experts, followed by a set of design goals derived from the requirements. We then detail our visualization techniques.

5.1 User Requirements

In this study, we have collaborated with a group of domain experts in communication and media studies comprising one professor, four postgraduate students, and one undergraduate student to understand the spatio-temporal diffusion of information on social media. After multiple brainstorming sessions, we have identified a set of user requirements and research questions for analyzing and exploring the dynamics of spatio-temporal diffusion of information regarding an event on Twitter.

R1. How will the overall spatial diffusion network evolve over time? When will significant changes occur in this spatial diffusion process?
R2. How do various geolocations across the world differ from each other on their roles in information diffusion on social media? Specifically, users in which geolocations will initiate, improve, or impede the information diffusion on social media?
R3. How will the roles of different geolocations in information diffusion evolve over time? When will a specific geolocation contribute most or least to the spatial diffusion of information? When will a specific geolocation exert the greatest influence on or receive the greatest influence from other geolocations?
R4. Given an identified spatial or temporal diffusion pattern, could plausible explanations or preliminary hypotheses be formed on possible events? For instance, when a group of geolocations are found to display unusual diffusion patterns in most time periods, is this phenomenon mainly caused by the geographical distance, cultural proximity, or/and linguistic similarity? Will the unusual patterns hold in various events?
R5. What hashtags are more likely to go viral across the globe (or across a group of selected geolocations)? When did they go viral? How can hashtags differ from each other in the properties of their spatial diffusion at a given time?

These user requirements help us shape the principles of visual design, and draw the roadmap for our visual system.

5.2 Design Goals

We derive the following design principles to design visualization techniques that help address the research questions.

G1. **Provide an overview of spatio-temporal diffusion.** The system should provide an overview of spatio-temporal diffusion to address the questions in R1. In particular, it should visually summarize the overall evolution of spatial diffusion network over time to help detect interesting and critical time points for further analysis (e.g., the time when the diffusion network concentrated in certain geolocations). The overall visualization also serves as a basis for analysts to drill in to identify detailed temporal or spatial diffusion patterns for in-depth exploration and analysis (R2–R5).

G2. **Unfold diffusion patterns.** When interesting overall diffusion patterns emerge, the users should be allowed to interactively see and examine the detailed diffusion patterns as well as context information regarding the patterns to enable in-depth analysis. SocialWave uses a map visualization to unfold any pattern discovered in the overview. It can distinctly convey the pair-wise diffusion among geolocations on a map at a given time period, which is important to address R2, R3, and R5. Moreover, temporal diffusion pattern between specific geolocations must be unfolded to users to support further analysis (R2 and R3).

G3. **Support comparative analysis.** The system should allow users to create a series of small coordinated multiples that show spatial diffusion among geolocations at different time periods to facilitate comparative analysis. Tracking and comparing spatial diffusion patterns at different periods to understand the commonalities and differences over time is particularly useful for R1, R2, R4, and R5. Different geolocations has different magnitude of impact, our visual design should provide multi-scale information to enable in-depth comparison.

G4. **Design intuitive visual representations.** Our collaborators have little experience in using advanced visualization systems. They prefer simple and intuitive visualizations to perform investigative tasks (R1–R5). Therefore, SocialWave combines a circular cartogram and a node-link diagram, which are easy to understand in a novel way to visually represent the diffusion among geolocations. Further, the design should be helpful in preserving user's geographical mental map.

G5. **Reduce visual clutter.** SocialWave visualization can be severely cluttered with an increasing number of nodes and edges of varying sizes in the diffusion network. The issue can easily degrade task performance or lead to misleading information (R1–R5). SocialWave seeks to effectively reduce visual clutter through a new layout algorithm that iteratively and judiciously bundles the edges and adjusts the node positions to remove overlaps while preserving the relative spatial positions of the nodes.

5.3 Visualization Techniques

This section introduces our visualization techniques. Figure 2 shows our user interface that has two main views: a temporal visualization (Figure 2(a)) for showing the overall trend of spatial diffusion among all geolocations over time, and a spatial visualization (Figure 2(b)) for displaying diffusion of information at a given time period. Figure 2(c) shows a hashtag view for selecting hashtag(s) of interest to be investigated. A history view for a user to save interesting findings is also provided on the right of the user interface (Figure 2(d)).

5.3.1 Spatial Visualization. **Visual Encoding.** In our design, we need to demonstrate the spatial diffusion among different places at a given time (G2), which could be transformed into a problem of visualizing a weighted graph with nodes and edges of varying sizes. We design a spatial visualization to address this problem by integrating a widely used node-link diagram into a well-known circular cartogram. By using the circular cartogram, we can clearly and intuitively demonstrate an attribute by the size of a circle, and do not need to compare complex shapes that represent different places (G4 and G5) (Dorling 1996). Other designs such as conventional choropleth or map distortion may lead to biased perception (Dorling 1995). Further, circular cartogram could preserve the approximate centroid of corresponding geolocations to maintain users' geographical mental map. The diffusion network is visualized with node-link diagram, which is widely used in conveying network structure intuitively (Figure 2(b)). The nodes in SocialWave represent places, and the node positions imply the geographical positions of the places. The size of a node

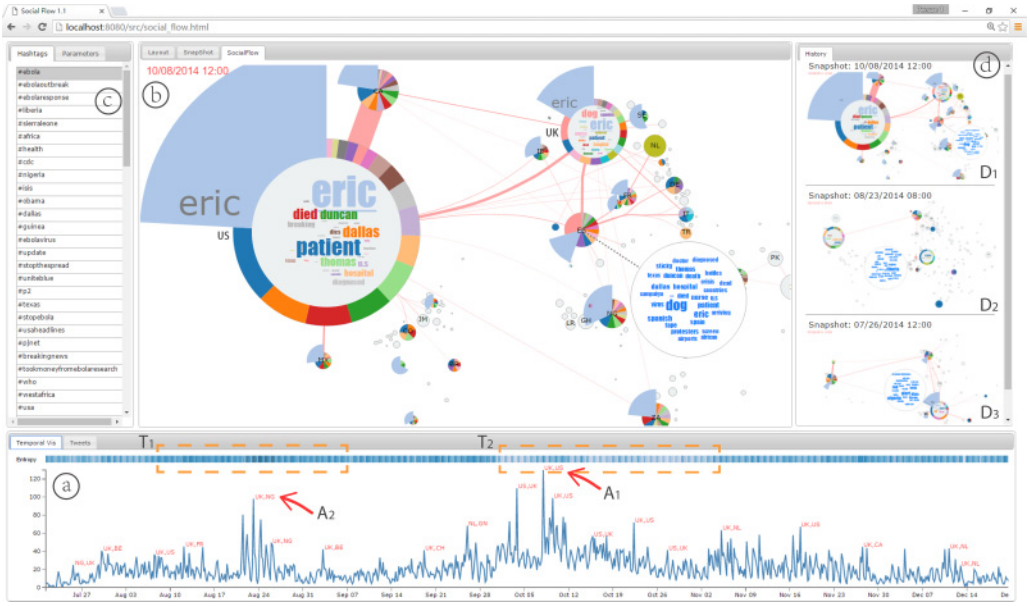

Fig. 2. Spatio-temporal diffusion of information on social media during outbreak of 2014 Ebola Epidemics with SocialWave. (a) The overall temporal trend of spatial diffusion among all geolocations over time; (b) spatial visualization for displaying diffusion of information at a given time period; (c) hashtag view for selecting hashtag(s) of interest to be investigated; (d) history view for comparing spatial diffusion patterns found at different time.

encodes the salience of the corresponding geolocation with respect to certain hashtag(s) of interest. The diffusion among geolocations is represented by edges, and the width of edges encodes the diffusion power.

To provide a more detailed overview of the spatial visualization, we propose a multi-scale node visualization (G1, G3, and G4). According to the size of the nodes, three categories of nodes are provided in the SocialWave, namely, *word node*, *pie node*, and *plain node*. In a word node (see the nodes representing US and UK in Figure 2), salient keywords are embedded in the node to distinctly and intuitively summarize the conversations in the specific location. A word node is further surrounded by a donut chart to intuitively encode the proportion of different keywords, and the arc color corresponds to the color of the keywords in the node. For smaller nodes (i.e., pie nodes), we replace the nodes with pie charts (see the nodes representing ES, FR, and NG, and so on, in Figure 2). Each slice represents one keyword, and the color helps distinguish different keywords. The size of each slice in a pie node is proportional to salience of corresponding keywords. To reduce visual clutter (G5), we decide to leave the nodes that are much smaller as it is (plain nodes). The slices of the donut chart and pie chart are ordered anti-clockwise. The color is globally consistent in one spatial diffusion network,

Spatial Layout Generation. Visualization of a diffusion network can be viewed as a problem of visualizing a weighted graph in a limited screen, while the nodes occupy initial positions and the size of nodes and width of edges may vary to different extents. Simply combining both the node-link diagram and circular cartogram may produce unappealing and ineffective results. For example, severe overlap and clutter can be introduced, because the nodes and edges are of different sizes (Figure 3(a)). This situation can hinder users in exploring diffusion patterns. Moreover, the

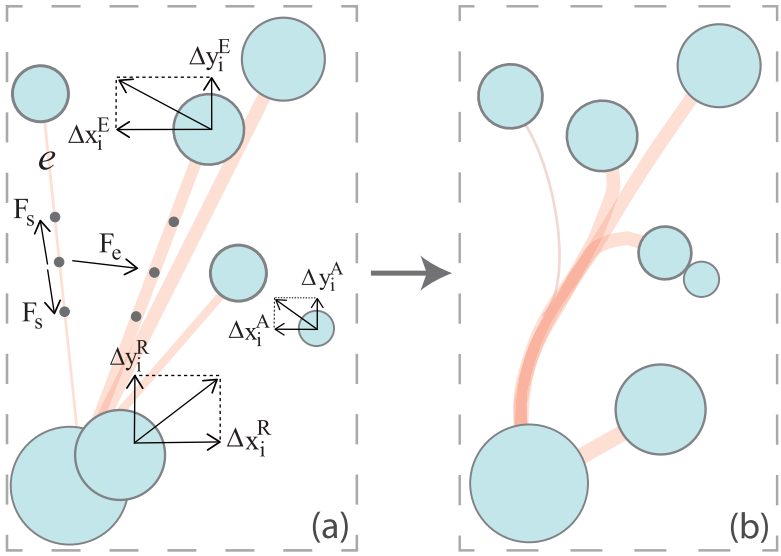

Fig. 3. Illustration of the algorithm used to rearrange the nodes. (a) Initial layout without nodes displacement and edge bundling; (b) rearranged layout without overlap among nodes and edges.

crossing among edges will further aggravate the visual clutter. The former issue could be handled by scaling the nodes size and edge width. However, the overlap between edges and nodes may still exist and some nodes and edges can be too small or too thin to be perceived. The latter issue may be addressed by edge bundling methods, but the bundled edges may still overlap with other nodes nearby.

Alternatively, the visualization layout problem can be directly transformed into a non-linear optimization problem with various types of non-linear constraints, such as avoiding overlaps among nodes, avoiding overlaps among nodes and edges, and preserving relative positions of nodes. However, the number of constraints can increase quickly as the number of nodes increases and the graph becomes cluttered. Thus, the optimization in a cluttered graph with respect to many constraints can easily result in large search space with many local minimums. We implemented a few methods by formulating the problems into different optimization problems, such as least squares optimization and general non-linear optimization. After several experiments, we found it is difficult, if not impossible, to find a good solution using the methods in such huge search space. Most optimization results still suffer from the overlap problem of edges and nodes. Furthermore, interactive performance can be hardly achieved.

To reduce clutter and create a distinct and occlusion-free graph visualization, we propose a new layout algorithm to iteratively rearrange the nodes and displace the edges with repulsive and attractive forces exerted on the nodes and edges in the graph. We classify the exertion of repulsive and attractive forces on the nodes and edges into four categories: *repulsive forces among overlapped nodes*, *repulsive forces among overlapping nodes and edges*, *attractive forces among geographical neighboring nodes*, and *repulsive and attractive forces among compatible edges*.

Repulsive Forces among Overlapped Nodes. In general, nodes that overlap with each other should be repelled outward to remove occlusion. Directly applying the repulsive force function proposed in circular cartogram (Dorling 1996) will pose two major challenges. First, some overlapping nodes are connected with edges, but the nodes shifted by the force will be in close contact

with neighboring nodes. Thus, the connection between the nodes cannot be well revealed, because they are too close. Therefore, extra space needs to be reserved to display the edges legitimately. The second challenge lies on the size of the nodes. A large node may overlap severely with its surrounding nodes, and it can be repelled by those nodes. Given the preattentive effect and visual salience of the node size (Ware 2012), a dramatic displacement of the large nodes can break users' geographical mental map. Thus, a larger node should be shifted less. On the basis of these concerns, we propose a new function for computing the repulsive force exerted by node n_j to its adjacent overlapping node n_i with respect to the x and y dimensions as follows:

$$\Delta x_i^R = w_{i,j}\frac{O_{i,j}(x_j - x_i)}{S_i||n_i - n_j||}, \quad \Delta y_i^R = w_{i,j}\frac{O_{i,j}(y_j - y_i)}{S_i||n_i - n_j||}, \tag{4}$$

where $w_{i,j}$ represents the weight of the edge connecting node n_i and n_j, $O_{i,j}$ is the overlap size between node n_i and n_j, and S_i is the area of node n_i. This repulsive force exerted on a node is illustrated in Figure 3(a) as notations Δx_i^R and Δy_i^R. We compute $O_{i,j}$ by $r_i + r_j - ||n_i - n_j||$, where r_i and r_j represent the radius of node n_i and n_j, and $||n_i - n_j||$ is the distance between n_i and n_j. The repulsive force is enabled only when $O_{i,j}$ is larger than zero. In Equation (4), $w_{i,j}$ aims to increase the repulsive force exerted on node n_i, and S_i is used to weigh down the force.

Repulsive Forces among Overlapped Nodes and Edges. In addition to the overlap issue among nodes, overlaps may also exist between nodes and edges. Thus, the nodes should be repelled to avoid occlusion with the edges. We denote the overlap between edge e_k and node n_i as $O_{k,i}$, and the centroid of the overlap area as n_c. We formulate the repulsive force exerted from edge e_k on node n_i as follows:

$$\Delta x_i^E = \omega_{k,i}\frac{O_{k,i}(x_c - x_i)}{S_i||n_c - n_i||}, \quad \Delta y_i^E = \omega_{k,i}\frac{O_{k,i}(y_c - y_i)}{S_i||n_c - n_i||}. \tag{5}$$

This repulsive force exerted on a node is illustrated in Figure 3(a) as notations Δx_i^E and Δy_i^E. The repulsive force from an edge to a node is not only proportional to the size of the overlap between them, but also proportional to the distance between the edge and center of the node. For instance, if a thin edge lies near the center of a node, the repulsive force will be small because of the small overlap. This situation will result in a large number of iterations for final visualization generation. Thus, we introduce parameter $\omega_{k,i}$ to handle this issue. We compute $\omega_{k,i}$ by $r_i/||n_c - n_i||$. The parameter helps enhance the repulsive force when a thin edge lies in a large node, which accelerates the iteration process. A closer edge to the center of a node corresponds to a larger repulsive force exerted on the node from the edge.

Attractive Forces between Neighboring Nodes. To preserve users' geographical mental map, the displaced nodes should be close to their original geographical neighboring nodes. Thus, we expect that the originally neighboring nodes should be attracted to each other. We formulate the attractive force exerted from node n_j on neighboring node n_i as follows:

$$\Delta x_i^A = \frac{M_{i,j}(x_j - x_i)}{w_{i,j}S_i||n_i - n_j||}, \quad \Delta y_i^A = \frac{M_{i,j}(y_j - y_i)}{w_{i,j}S_i||n_i - n_j||}, \tag{6}$$

where $M_{i,j}$ is the distance between node i and node j, which is defined as $M_{i,j} = ||n_i - n_j|| - r_i - r_j$. The attractive force is valid only when $M_{i,j}$ is larger than 0. We use $w_{i,j}$ in the denominator to prevent the connected nodes from getting too close to each other, because extra space should be maintained for displaying the edges. The attractive force exerted on a node is illustrated in

Figure 3(a) as notations Δx_i^A and Δy_i^A. The geographical neighbors of a geolocation are extracted from the GeoNames database.[4]

Finally, the total repulsive and attractive forces exerted on node i can be derived as follows:

$$\Delta x_i = \beta(\Delta x_i^R + \Delta x_i^E) + (1-\beta)\Delta x_i^A, \tag{7a}$$

$$\Delta y_i = \beta(\Delta y_i^R + \Delta y_i^E) + (1-\beta)\Delta y_i^A, \tag{7b}$$

where β is the ratio controlling the proportion between the repulsive and attractive forces. The bigger the β, the larger the repulsive force is, and the faster the iteration ends. However, with a larger repulsive force, the nodes may be pushed outward too far, thus leading to an unappealing result. We set $\beta = 0.9$ after experiments to create reliable and appealing layouts.

Repulsive and Attractive Forces between Edges. To further reduce visual clutter, the edges representing the diffusion between different geolocations are bundled to improve graph readability. One straightforward solution is to bundle the edges in an occlusion-free graph layout created by a layout algorithm by using only the three aforementioned forces. However, this method can easily introduce new overlaps between the nodes and bundled edges. Therefore, the edges should be bundled simultaneously with the process of the nodes displacement. We adopt a force-directed edge bundling method (Holten 2006) to aggregate compatible edges in each iteration. Compatible edges are the edges that have similar angles, similar lengths, similar positions, and so on (Holten 2006). As illustrated in Figure 3(a), a subdivision point on edge e is attracted by the subdivision points on the edges compatible to e in the context of electrostatic force (F_e). The point is also attracted by nearby subdivision points on e in the context of spring force (F_s).

A visualization example generated by the layout algorithm is demonstrated in Figure 2(b). Each node represents a country, and the diffusion among countries is represented by edges. We can see that the displaced nodes have no overlap with other nodes and bundled edges, while relative position among the nodes are preserved as much as possible.

5.3.2 Temporal Visualization. **Visual Encoding.** To reveal the overall temporal trend of spatial diffusion (G1), we adopt a line chart (Figure 2(a)) to present the time-varying graph centrality of the weighted diffusion graph. The line chart can reveal the dynamic variation of the degree of centralization of the diffusion network (Freeman 1979) and identify when the diffusion spreads from one or most influential geolocations. A larger graph centrality leads to the stronger centralization of the diffusion network. The texts near various peaks on the line chart indicate abbreviation names of the most influential geolocations.

We also provide a 1D heatmap to present the temporal variation of spatial property of a hashtag(s) (Figure 2(a)). We use spatial entropy that is defined in Kamath et al. (2013) as the property to illustrate how the distribution of a hashtag(s) evolves over time. The color of each bar in the 1D heatmap represents the spatial entropy of a hashtag(s) at each time. Darker color indicates higher spatial entropy, which implies that the hashtag(s) are evenly distributed across the world, and vice versa.

Temporal Layout Generation. We calculate the graph centrality of the spatial diffusion network at a given time with the method proposed in Freeman (1979). The centrality of a node (e.g., the degree centrality, betweenness centrality) must be first calculated. Given that the edges in the diffusion network have different weights, we use weighted degree centrality, which accounts for both number of edges associated with a node and the weights of the edges, as the centrality measurement of a node n_i. The weighted degree centrality of node n_i is defined as $c(n_i) = \sum_{j \in N(n_i)} w_{i,j}$.

[4] http://www.geonames.org/.

$N(n_i)$ is the set of neighboring nodes connected with node n_i, and $w_{i,j}$ is the weight of the edge associating node n_i and n_j. With the weighted degree centrality of each node, we calculate the graph centrality of diffusion network at time t as follows:

$$C^t = \frac{\sum_{i=1}^{n}(c_*(n_i) - c(n_i))}{max \sum_{i=1}^{n}(c_*(n_i) - c(n_i))}, \tag{8}$$

where $max \sum_{i=1}^{n}(c_*(n_i) - c(n_i))$ represents the maximum possible sum of differences of node centrality for any graph of n points, and $c_*(n_i)$ is the largest weighted degree centrality in the graph. With degree centrality as the measurement of node centrality, the maximum sum of differences in the denominator can be determined by $(n-2)(n-1)$ with respect to a star graph (Freeman 1979).

5.3.3 User Interactions. SocialWave supports various basic and advanced interactions to address different analytical tasks (G2 and G3).

Overview First and Details-on-Demand. SocialWave provides a summary of the spatio-temporal diffusion patterns to assist users in finding interesting and critical time points. Multiple perspectives of detail are available in SocialWave to help reveal the detailed information of the diffusion. Users can examine the detailed spatial diffusion pattern by directly clicking on the line chart at corresponding time. In the spatial visualization, users can further investigate the detailed temporal diffusion pattern between locations of interest by clicking the edge representing the diffusion.

Comparative Analysis Support. Users could hover over a keyword, a pie slice, or a donut slice to highlight the proportion of the keyword in different locations (see the pop-up slices representing "eric" in Figure 2). When users find an interesting spatial diffusion pattern, or a pattern together with word cloud, they could save the pattern to a history view to form small multiples for further comparison. Users are allowed to navigate in the history view and restore corresponding patterns. They are also allowed to investigate the pair-wise diffusion power between two geolocations by hovering over one edge. A green circle will emerge on the edge, and the position of the circle implies the proportionality of the diffusion power. The more distant from the circle to a geolocation, the large diffusion power exerted from the geolocation to the other.

Detailed conversations examination. Users can click on a node, and a detailed word cloud will pop up to present a visual summary of the tweets that are posted in the location represented by the node at that time period. The word cloud facilitates the identification of significant keywords. Users can further click on a keyword of interest to examine the detailed tweets containing the keyword, as well as the users in the same place and at the same time.

6 EVALUATION

This section presents two case studies and a user evaluation to demonstrate the usability and effectiveness of SocialWave.

6.1 Case Studies

We collected two large-scale Twitter datasets about two events that massively trended on social media in 2014, namely, Ebola Epidemics and Ferguson Unrest, to test the effectiveness and usefulness of SocialWave. The Ebola Epidemics dataset contains 17,125,091 tweets and 3,280,193 users ranging from July 20, 2014 to December 25, 2014; and the Ferguson Unrest dataset contains 22,476,117 tweets and 2,586,423 users ranging from August 5, 2014 to December 12, 2014.

Spatio-temporal Diffusion in Ebola Epidemics. This case study is used to demonstrate how SocialWave helps users explore the spatio-temporal diffusion during the outbreak of Ebola Epidemics. We select #Ebola as the main propagation unit to explore the diffusion.

SocialWave allows for the quick identification of the overall temporal trend of spatial diffusion of information. Figure 2(a) illustrates a 1D heatmap and a line chart for showing the temporal variation of spatial entropy and centralization of the diffusion network with respect to #Ebola, respectively. The 1D heatmap presents the evolution of spatial distribution of #Ebola over time. The bars in time period T1 have darker color than those in time period T2 (highlighted within the orange dashed rectangles), indicating that #Ebola has a higher spatial entropy during T1. This finding implies that #Ebola was evenly distributed among different geolocations all over the world in T1, and was concentrated in certain geolocations in T2. The line chart below reveals when significant changes emerge in the spatial diffusion process as well as relevant influential geolocations. The texts near peaks indicate the most influential geolocations, and examining the texts reveals that the United States (US) and the United Kingdom (UK) occurred most frequently (see the Arrow A1 in Figure 2(a)). This pattern is more apparent in the period from October 5, 2014 to October 26, 2014, which approximately coincides with time period T2. This finding suggests that occurrence of #Ebola concentrated in US and UK, and that they initiated to diffuse information to other geolocations during the period.

SocialWave enables users to unfold detailed spatial diffusion patterns among different geolocations. We first focus on the time period with the highest graph centrality by directly clicking the peak that occurred on October 8, 2014 on the line chart. A spatial visualization (Figure 2(b)) is presented to demonstrate the spatial diffusion pattern during that period. Given the varying sizes of the nodes, we could immediately determine that US and UK are the geolocations most responsible for emergence of #Ebola on Twitter. These countries emitted plenty of edges with varying widths, revealing that US and UK played major roles in disseminating information with different diffusion power with respect to #Ebola at that time period. For instance, the edge connecting Canada (CA) and US is the widest among all of the edges. These patterns may likely be due to the occurrence of important events happened in these countries. Thus, we investigate the detailed information generated in these three countries (US, CA, and UK) by examining the word clouds, donut charts, and pie charts nested in the corresponding nodes (Figure 2(b)). We can clearly observe for both US and CA, the largest keyword in the word clouds is "eric," followed by other keywords such as "died" and "patient." Tweets containing the keyword "eric" are examined by directly clicking the keyword in the two word clouds, and we find that majority of the tweets in these two countries are reporting the death of the first person diagnosed with Ebola in the US (e.g., "*@Wall Street Journal: Thomas Eric Duncan, the first person to be diagnosed with Ebola in the U.S., has died, according to hospital*"). We speculate that the remarkable diffusion power between these two countries results from the linguistic similarity of the tweets and the close geographic distance between them.

In addition, we observe a number of edges connecting US and the European countries (e.g., UK, Spain (ES) and France(FR)) in Figure 2(b), even though these countries are of great distances from US. For further investigation, we select UK as our focus country because of the notable emergence of #Ebola in this country, as indicated by the size of the node and relatively stronger diffusion power between US and UK compared with that in other countries linked to US. Similarly, we find the keyword "eric" is also salient in UK. This finding can be considered proof of the diffusion phenomenon between US and UK that results from the comparable linguistic similarity despite the great distance between them. We also find other salient keywords such as "dog" and "nurse" in the word clouds. Review of the related tweets containing "dog" and "nurse" reveals that users in UK were tweeting and protesting about a Spanish court order about putting down a pet dog without quarantine (e.g., "*@The Telegraph: Thousands sign petition to save #Ebola dog as protesters clash with police*"). The owner of the dog, a nurse, contracted Ebola while treating Ebola patients. After examining the word clouds and raw tweets of nearby countries such as ES and FR, we observe similar patterns for this event (see the pop-up word cloud in Figure 2(b)). This event can be considered

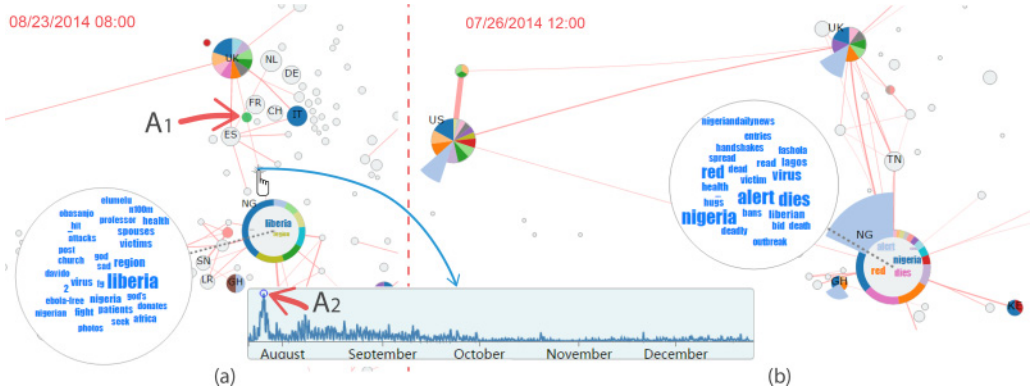

Fig. 4. (a) Visual summary of the spatial diffusion patterns in Ebola Epidemics dataset among different geolocations on August 23, 2014. (b) Spatial diffusion patterns found on August 23, 2014, and Nigeria (NG) served as the center for information diffusion on social media.

the reason for the information diffusion indicated by the edges among these countries (i.e., UK, ES, and FR) at that time. Moreover, in view of the differences in language of these countries, we speculate that the close distance among these countries served as an essential role in shaping the information diffusion among them at that time.

SocialWave can unfold the pair-wise and temporal diffusion patterns among geolocations of interest. Figure 4(a) shows the spatial diffusion pattern of another interesting time point (August 23, 2014), during which the graph centrality of the diffusion network reached another peak and the diffusion was centralized in UK and Nigeria (NG) (see Arrow A2 in Figure 2(a)). We can see that most of the edges spread out from UK and Nigeria. Thus, we examine why Nigeria is one of the centers for information diffusion and to what extent it influenced other countries during that time. We hover over the edges linking Nigeria, for instance, the edge between Nigeria and UK. The nodes representing Nigeria and UK are highlighted. A green circle emerges on the edge (see Arrow A1 in Figure 4(a)), and the distance from the position of the circle to Nigeria is larger than that to UK, implying that the users from Nigeria exerted more influence than the users from UK.

We then examine the tweets posted in Nigeria and find the keyword "Liberia" is the most salient keyword in the word cloud. After examining the tweets containing "Liberia," we find that the users in Nigeria were heavily tweeting about the imminent threat of Ebola in West Africa. We speculate that such massive amount of discussion about the Ebola threat resulted in the high diffusion power from Nigeria not only to nearby countries but also to distant countries such as UK. Our collaborating expert is also interested in investigating when Nigeria had the most intense interaction with UK. Thus, we click on the edge linking Nigeria and UK, and a line chart appears (Figure 4(a)), which demonstrates the temporal variation of the diffusion power between the two countries. We could see the diffusion power between Nigeria and UK reached its peak at the beginning (see Arrow A2 in Figure 4(a)). Clicking the peak shows the spatial diffusion at that time. We could obviously observe that Nigeria acted as a center of information diffusion across the world (Figure 4(b)). We then examine the word clouds of the tweets posted in Nigeria at that time, and find that Nigeria is as prominent as the keywords "alert" and "dies." Further review of the tweets containing the keywords shows that this pattern is related to the first death that occurred in Nigeria, which led to the panic for the Ebola infection.

SocialWave allows users to compare the detailed spatial diffusion in different time periods. Before switching to other time periods, users can save current spatial diffusion patterns as well as

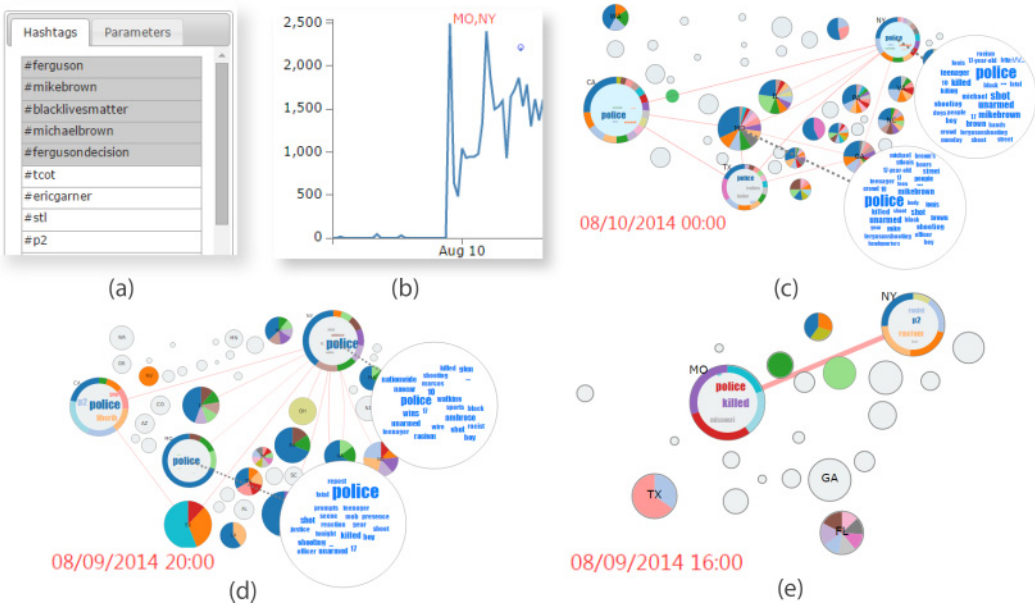

Fig. 5. Spatio-temporal diffusion patterns found in Ferguson Unrest dataset in August, 2014. Spatial diffusion patterns in three adjacent time windows (c, d, and e) are examined for investigation of the temporal patterns of information diffusion in a short period.

the word clouds in the history view to form small multiples for further comparative analysis. Figure 2(d) shows three small multiples of three time periods, namely, October 8, 2014 (D1), August 23, 2014 (D2), and July 26, 2014 (D3), which are the same time periods investigated above. As time went on, the size of the node representing US increases rapidly, indicating that the salience of #Ebola in US has become much more significant, and US has started to act as the center of the information diffusion on social media with respect to #Ebola. Moreover, the small multiples show that the number of countries involved have increased and that the size of most nodes have been getting larger, indicating that #Ebola has formed a global phenomenon. This pattern demonstrates that small multiples is useful in providing a different perspective for investigating spatial diffusion patterns.

Spatio-temporal Diffusion in Ferguson Unrest. The scalability of SocialWave is demonstrated by the analysis of the spatio-temporal diffusion of information regarding Ferguson Unrest. Five hashtags that massively trended at that time, shown in Figure 5(a), are selected to investigate the spatio-temopral diffusion.

With the use of SocialWave, users can discover how information is diffused in a short time when an unforeseen event occurs. The line chart for the temporal variation of diffusion network centrality (Figure 5(b)) instantly reveals a significant peak at the beginning. By clicking the peak, we are navigated to the visualization of the spatial diffusion of a possible breaking event, which occurred on August 10, 2014 (Figure 5(c)). The emitted edges from the nodes indicates that Missouri (MO), New York (NY), and California (CA) are responsible for the significant peak and act as the centers for information diffusion. An examination of the word clouds for the states reveals that the keyword "police" was very salient. A further investigation of the tweets containing "police" shows that this pattern was related to the protest against the police that was caused by the gun shooting

down of a 17-year-old boy by the police in Ferguson, MO. However, our collaborating expert was more interested in how such unforeseen event diffused on social media at the very moment that it was happening.

In SocialWave, users can simply press the "Left" key to explore the spatial diffusion in previous time windows. Figures 5(d) and 5(e) show the spatial diffusion patterns of two previous time windows before the current one, that is, 4 and 8h beforehand. When examining the spatial diffusion in the first time window (Figure 5(d)), many keywords (e.g., "police" and "unarmed") related to the event can be observed in the word clouds of the previous states, and that the keywords are more salient in Missouri than in other states. Performing the same evaluations in the second time window (Figure 5(e)), we find that only a few keywords (e.g., "killed" and "police") are related to the event in the word clouds of Missouri, and we fail to find relevant keywords in the word clouds of New York. This temporal pattern implies that the users in New York were presumably influenced by the users in Missouri with a certain time lag. We could further hypothesize that recency effect influenced how this unforeseen event was diffused from Missouri to other geolocations at that time. The spatial diffusion in each time period can be snapshotted in the history view to facilitate convenient side-by-side comparisons. This case study shows that SocialWave enables users to detect and analyze the differences in spatial diffusion within adjacent time periods.

6.2 User Feedback

To evaluate the effectiveness of SocialWave, we interviewed one professor who engaged in communication research for more than 10 years, as well as four postgraduate students and one undergraduate student who are majoring in communication/news media. We first described the visual encoding and user interactions in SocialWave then demonstrated the patterns that were observed in the case studies. Their feedback is summarized as follows.

Model Design. The proposed model received positive feedback during the interview from the experts. They all confirmed that the model can well capture the dynamics of the information diffusion, and estimating the parameters using real world flight network dataset is practical, since the ground-truth data with respect to information diffusion is not easy to quantify and collect. Regarding the improvement of the proposed model, one expert added that "current model is based on the node-level (i.e., the variables on both side of the Equation (1) represent the salience of information of a region), the model could be improved by adding an auxiliary model that takes the dyadic-level of the diffusion network into count." For example, the dependent variable of the equation could be the connection strength of two nodes. Regarding the regression feature of our model, the expert suggested that splitting current model into a multi-level model could derive more information for further analysis and following visualization. For example, the salience of a hashtag in location i at time t could be related to a function of various factors including salience of the hashtag in location i at time t-1, salience of the hashtag in location j at time t-1, and semantic distance between location i and j at time t-1. The coefficient before each factor could be related to a function of various time-invariant factors including geographical distance, culture similarity, and flight flow among location i and j.

Visual Design. All the users agreed that exploring spatial diffusion patterns among different geolocations is not hindered by the displacement of the nodes that represent the geolocations. They also agreed that despite being simple, the spatial visualization can intuitively convey spatial diffusion. The professor particularly liked the idea of placing spatial diffusion patterns from different time periods in the history view. He added that "*such component is convenient and valuable for me to relate and compare current spatial diffusion pattern with previous observed patterns.*" The green circle placed on the edge was also appreciated by the professor, who commented that: "*The design is so cute and it is brilliant to re-use an existing visual element to present the dyadic diffusion*

pattern." However, two of the participants had difficulty in relating the degree of diffusion network centrality presented on the line chart to the detailed spatial diffusion pattern, because they were unfamiliar with the concept of graph centrality. Regarding the improvement of the visual design, one expert suggested that placing the names of the most influential geolocations near each peak on the graph centrality line chart should be useful for quickly identifying the overall trend of the roles played by the geolocations.

Usability. All the users confirmed the usefulness and effectiveness of the system, and they all expressed their eagerness for the system to be put online. The professor mentioned that the process for exploring patterns in such way of "from overview to detail" is reasonable, and that the user interactions were smooth and easy to master. All of the students appreciated the word clouds and the coordinated list view of tweets for investigating hidden insights behind the observed patterns. Regarding the improvement of usability, three students suggested that the system should support exporting the analysis data, such as numerical diffusion power and the keyword frequency in the word clouds, to excel files.

7 DISCUSSION

The visual analytic system, SocialWave, can provide communication researchers with an interactive, informative, and insightful platform to explore the spatio-temporal characteristics of information diffusion, which can lay a solid foundation for more sophisticated explanatory and predictive analysis.

Second, the visual analytic system is of great practical values in different applications. The SocialWave can serve as a competent tool for commercial organizations to track the diffusion of innovative products/ideas across the world, which can help business leaders make well-informed decisions. In public relations (PR) and advertising, the SocialWave can provide valuable and accurate geographical and temporal information for PR/advertising practitioners in designing, implementing, and assessing viral marketing campaigns on social media.

The proposed visual design strictly follows a participatory design process (from the start of this project to final evaluation) under a close cooperation with domain experts, who prefer simple and easy-to-use visualizations. Although the visualization is intuitive, the underlying techniques are non-trivial and grounded on a fundamental theory (i.e., the graph centrality in the temporal visualization) and solid optimization (i.e., the layout generation in the spatial visualization). Our graph layout technique could serve as a prompt for an open thread and induce others to come forward with their solutions with respect to following challenges in our complex graph layout case: a dynamic directed weighted graph with initial positions of nodes, and the nodes and edges are of different sizes. The temporal visualization visually summarizes structural changes in diffusion network centrality over time, such that a user can quickly identify interesting time periods. This top-down visualization strategy works effectively for exploring the spatio-temporal patterns. However, as suggested by domain experts, there are still scenarios where they want to find and see a certain diffusion pattern, which can be specified by various conditions. We plan to investigate this issue and design an intuitive user interface to support this task.

Limitation. Our work still has some limitations to overcome. Regarding the social gravity model, it is challenging to perform cross-validation, since the ground truth datasets regarding the spatial interaction among locations with different spatial and temporal granularities are not easy to quantify and collect. Second, the ground truth of spatial diffusion strength among locations regarding different events in different time periods may vary significantly. In the future work, we plan to further evaluate our model against more real world dataset such as international trading data and immigration data, and the importance of different model factors and their interactions, which are critical for us to learn more about the behavior of the model. For example, what will

the model perform without the cultural proximity? Does this parameter affect the model much? Regarding the node-link diagram to demonstrate the data distribution, it is not as intuitive as traditional geographical views, since the positions of the nodes on the diagram cannot reveal the geographical positions accurately. Currently, we provide users an option to switch between node-link view and geographical view. We plan to further explore potential visualizations that could achieve an appropriate trade-off between these two issues.

8 CONCLUSION

In this study, we present SocialWave, a visual analytics system that couples an advanced diffusion model and interactive visualization to explore and analyze spatio-temporal diffusion of information on social media. The diffusion model is extended from a classic gravity model by considering four factors, namely, geographic distance, recency effect, cultural proximity, and linguistic similarity, which have been regarded as theoretically significant factors in communication and media studies. Our proposed visualizations comprise two main visualizations: the temporal and spatial visualizations, to visualize the spatio-temporal patterns measured by the diffusion model. The temporal visualization displays the trend of changes in diffusion network centralization over time, from which a user can quickly locate important time periods. To visualize the spatial diffusion within a given time period, the spatial visualization uses a novel layout algorithm to create an occlusion-free, expressive visualization with an integration of a multi-scale catogram and a weighted diffusion network. In the future, we plan to explore how to extend the model and data processing steps to incorporate unsupervised learning techniques to increase likelihood of finding unknown patterns, and enhance the visualizations to handle streaming data.

REFERENCES

James E. Anderson and Eric van Wincoop. 2003. Gravity with gravitas: A solution to the border puzzle. *Amer. Econ. Rev.* 93, 1 (2003), 170–192. DOI: http://dx.doi.org/10.1257/000282803321455214

Harald Bosch, Dennis Thom, Florian Heimerl, Edwin Puttmann, Steffen Koch, Robert Kruger, Michael Worner, and Thomas Ertl. 2013. Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering. *IEEE Trans. Visual. Comput. Graph.* 19, 12 (2013), 2022–2031.

Frances Cairncross. 2001. *The Death of Distance: How the Communications Revolution is Changing Our Lives.* Harvard Business Press.

Nan Cao, Yu-Ru Lin, Xiaohua Sun, David Lazer, Shixia Liu, and Huamin Qu. 2012. Whisper: Tracing the spatiotemporal process of information diffusion in real time. *IEEE Trans. Visual. Comput. Graph.* 18, 12 (2012), 2649–2658.

James Caverlee, Zhiyuan Cheng, Daniel Z. Sui, and Krishna Yeswanth Kamath. 2013. Towards geo-social intelligence: Mining, analyzing, and leveraging geospatial footprints in social media. *IEEE Data Eng. Bull.* 36, 3 (2013), 33–41.

Justin Cheng, Lada Adamic, P. Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted? In *Proceedings of the International Conference on World Wide Web*. 925–936.

Weiwei Cui, Hong Zhou, Huamin Qu, Pak Chung Wong, and Xiaoming Li. 2008. Geometry-based edge clustering for graph visualization. *IEEE Trans. Visual. Comput. Graph.* 14, 6 (2008), 1277–1284.

Peter Sheridan Dodds, Eric M. Clark, Suma Desu, Morgan R. Frank, Andrew J. Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M. Kloumann, James P. Bagrow, et al. 2015. Human language reveals a universal positivity bias. *Proc. Natl. Acad. Sci. U.S.A.* 112, 8 (2015), 2389–2394.

Daniel Dorling. 1995. *A New Social Atlas of Britain.* John Wiley and Sons, Chichester, England.

Daniel Dorling. 1996. Area cartograms: Their use and creation. In *Proceedings of Concepts and Techniques in Modern Geography.*

Peter Eades, Qing-Wen Feng, and Xuemin Lin. 1997. Straight-line drawing algorithms for hierarchical graphs and clustered graphs. In *Proceedings of Graph Drawing*. 113–128.

David Easley and Jon Kleinberg. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World.* Cambridge University Press.

Linton C. Freeman. 1979. Centrality in social networks conceptual clarification. *Social Networks* 1, 3 (1979), 215–239.

The Guardian. 2011. Riot Rummors (2011). Retrieved from https://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-twitter.

Ivan Herman, Guy Melançon, and M. Scott Marshall. 2000. Graph visualization and navigation in information visualization: A survey. *IEEE Trans. Visual. Comput. Graph.* 6, 1 (2000), 24–43.

Chien-Tung Ho, Cheng-Te Li, and Shou-De Lin. 2011. Modeling and visualizing information propagation in a micro-blogging platform. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*. 328–335.

Geert Hoftede, Gert Jan Hofstede, and Michael Minkov. 2010. *Cultures and Organizations: Software of the Mind*. McGraw-Hill.

Danny Holten. 2006. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Trans. Visual. Comput. Graph.* 12, 5 (2006), 741–748.

Danny Holten and Jarke J. Van Wijk. 2009. Force-directed edge bundling for graph visualization. *Comput. Graph. Forum* 28, 3 (2009), 983–990.

Sergey Ioffe. 2010. Improved consistent sampling, weighted minhash and l1 sketching. In *Proceedings of IEEE International Conference on Data Mining*. 246–255.

Walter Isard and David F Bramhall. 1960. Gravity, potential and spatial interaction models. *Methods of Regional Analysis* (1960), 493–568.

F. Thomas Juster and Frank P. Stafford. 1991. The allocation of time: Empirical findings, behavioral models, and problems of measurement. *J. Econ. Lit.* 29, 2 (1991), 471–522.

Krishna Y. Kamath and James Caverlee. 2013. Spatio-temporal meme prediction: Learning what hashtags will be popular where. In *Proceedings of the ACM International Conference Information & Knowledge Management*. 1341–1350.

Krishna Y. Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. 2013. Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proceedings of the International Conference on World Wide Web*. 667–678.

Michael Kaufmann and Dorothea Wagner. 2001. *Drawing Graphs: Methods and Models*. Springer Science & Business Media.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 497–506.

Jure Leskovec and Christos Faloutsos. 2006. Sampling from large graphs. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 631–636.

David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. 2005. Geographic routing in social networks. *Proc. Natl. Acad. Sci. U.S.A.* 102, 33 (2005), 11623–11628.

Yu Liu, Zhengwei Sui, Chaogui Kang, and Yong Gao. 2014. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PloS One* 9, 1 (2014), e86026.

Sheng-Jie Luo, Chun-Liang Liu, Bing-Yu Chen, and Kwan-Liu Ma. 2012. Ambiguity-free edge-bundling for interactive graph visualization. *IEEE Trans. Visual. Comput. Graph.* 18, 5 (2012), 810–821.

Sofus A. Macskassy and Matthew Michelson. 2011. Why do people retweet? Anti-homophily wins the day! In *Proceedings of the International Conference on Weblogs and Social Media*. 209–216.

Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. 2011. Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 227–236.

Miller McPherson, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.* (2001), 415–444.

Tai-Quan Peng, Mengchen Liu, Yingcai Wu, and Shixia Liu. 2015. Follower-followee network, communication networks, and vote agreement of the US members of congress. *Commun. Res.* (2015), 0093650214559601.

Doantam Phan, Ling Xiao, Ron Yeh, and Pat Hanrahan. 2005. Flow map layout. In *Proceedings of the IEEE Symposium on Information Visualization*. 219–224.

Davood Rafiei. 2005. Effectively visualizing large networks through sampling. In *Proceedings of the IEEE Visualization*. 375–382.

Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. 2011. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the International Conference on World Wide Web*. 695–704.

Shahar Ronen, Bruno Gonçalves, Kevin Z. Hu, Alessandro Vespignani, Steven Pinker, and César A. Hidalgo. 2014. Links that speak: The global language network and its association with global fame. *Proc. Natl. Acad. Sci. U.S.A.* 111, 52 (2014), E5616–E5622.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the International Conference on World Wide Web*. 851–860.

Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, and Vito Latora. 2010. Distance matters: Geo-social metrics for online social networks. In *Proceedings of the International Conference on Online Social Networks*. 8–8.

David Selassie, Brandon Heller, and Jeffrey Heer. 2011. Divided edge bundling for directional network data. *IEEE Trans. Visual. Comput. Graph.* 17, 12 (2011), 2354–2363.

Ashish Sen and Tony E. Smith. 1995. *Gravity Models of Spatial Interaction Behavior*. Springer.

Paolo Sgrignoli, Rodolfo Metulini, Stefano Schiavo, and Massimo Riccaboni. 2015. The relation between global migration and trade networks. *Physica A: Stat. Mech. Appl.* 417 (2015), 245–260.

Moritz Stefaner. 2013. Revisit. (2013). http://moritz.stefaner.eu/projects/revisit/.

Guodao Sun, Yingcai Wu, Shixia Liu, Tai-Quan Peng, Jonathan J. H. Zhu, and Ronghua Liang. 2014. EvoRiver: Visual analysis of topic coopetition on social media. *IEEE Trans. Visual. Comput. Graph.* 20, 12 (2014), 1753–1762.

The New York Times. 2011. Project Cascade (2011). Retrieved from http://nytlabs.com/projects/cascade.html.

Peter Trudgill. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Lang. Soc.* 3, 2 (1974), 215–246.

Kevin Verbeek, Kevin Buchin, and Bettina Speckmann. 2011. Flow map layout via spiral trees. *IEEE Trans. Visual. Comput. Graph.* 17, 12 (2011), 2536–2544.

Cécile Viboud, Ottar N. Bjørnstad, David L. Smith, Lone Simonsen, Mark A. Miller, and Bryan T. Grenfell. 2006. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* 312, 5772 (2006), 447–451.

Fernanda Viégas, Martin Wattenberg, Jack Hebert, Geoffrey Borggaard, Alison Cichowlas, Jonathan Feinberg, Jon Orwant, and Christopher Wren. 2013. Google+ ripples: A native visualization of information flow. In *Proceedings of the International Conference on World Wide Web*. 1389–1398.

Colin Ware. 2012. *Information Visualization: Perception for Design (Interactive Technologies)* (3rd ed.). Morgan Kaufmann.

Yingcai Wu, Shixia Liu, kai Yan, Mengchen Liu, and Fangzhao Wu. 2014. Opinionflow: Visual analysis of opinion diffusion on social media. *IEEE Trans. Visual. Comput. Graph.* 20, 12 (2014), 1763–1772.

Jian Zhao, Nan Cao, Zhen Wen, Yale Song, Yu-Ru Lin, and Christopher Collins. 2014. # fluxflow: Visual analysis of anomalous information spreading on social media. *IEEE Trans. Visual. Comput. Graph.* 20, 12 (2014), 1773–1782.

Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban computing: Concepts, methodologies, and applications. *ACM Trans. Intell. Syst. Technol. (TIST)* 5, 3 (2014), 38.

Received December 2016; revised May 2017; accepted June 2017