SPATIOTEMPORAL ANALYSIS OF INFORMATION DIFFUSION IN ONLINE SOCIAL NETWORKS

by

Manqi Li
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Earth Systems and Geoinformation Sciences

Committee:

_____ Dr. Arie Croitoru, Dissertation Director

_____ Dr. Anthony Stefanidis, Committee Member

_____ Dr. Andrew Crooks, Committee Member

_____ Dr. Ruixin Yang, Committee Member

_____ Dr. Dieter Pfoser, Department Chairperson

_____ Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science

_____ Dr. Peggy Agouris, Dean, College of Science

Date: _____ Spring Semester 2019
George Mason University
Fairfax, VA

Spatiotemporal Analysis of Information Diffusion in Online Social Networks

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Manqi Li
Master of Science
State University of New York College of Environmental Science and Forestry, 2012
Bachelor of Science
Wuhan University, 2010

Director: Arie Croitoru, Professor
Department of Geography and GeoInformation Science

Spring Semester 2019
George Mason University
Fairfax, VA

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

SPATIOTEMPORAL ANALYSIS OF INFORMATION DIFFUSION IN ONLINE SOCIAL NETWORKS

Manqi Li, Ph.D.

George Mason University, 2019

Dissertation Director: Dr. Arie Croitoru

Understanding the dynamics of information diffusion in social networks contributes to a wide range of social studies. Among social networks, online social networks have drawn growing interest due to their richness, availability, and popularity nowadays. Such networks, which are often embedded in geographical space, have enabled information to spread at a relatively lower cost and higher speed and reach, compared to traditional ways of communication. This dissertation aims at exploring the spatiotemporal patterns of information diffusion in discussion about real-world events in online social networks, with special interest in geographical characteristics and representation. Specifically, this dissertation presents a methodology for studying and analyzing information diffusion in geographic space between sources and sinks of information. By doing so it highlights the information diffusion mechanisms that are in play at the intersection of the cyber and

geographical environment, which can provide additional insights for higher-level decisions making.

This dissertation also addresses the widely existing demand for traceable individual point information in data streams with geographical information, by designing an improved density-based stream clustering method. The method used not only meets the demand for finding cluster shapes, maintaining individual point information, and articulating point-cluster relationships, but also serves as the basis for spatiotemporal analysis and discovery of patterns hidden in the data.

Keywords: Information diffusion, Spatiotemporal analysis, Online social networks, Geographical data streams, Stream clustering

1 INTRODUCTION

Communication has been at the forefront of relationship building since prehistoric times. The way we communicate with each other has changed significantly over time with the advancement in technology. Changes in communication technology over the time have been revolutionary: from earlier radio and TV, to the emergence of the World Wide Web and Web 2.0 technology, and to the currently prevailing social media. The ways we communicate have therefore moved from physical to a blend of physical and cyber world, and have transformed from the typical one-on-one interpersonal interactions to the many-to-many interactive dialogues.

The recent advent of online and mobile social media services has enabled information to spread at relatively higher efficiency and lower cost, compared to traditional ways such as face-to-face contact. During recent years, increasing numbers of users on social media sites are detected. In 2018, about 68% U.S. adults get news on social media platforms, which was 49% in 2012 (Gottfried and Shearer, 2016; Matsa and Shearer, 2018). In light of this growing popularity, social media sites keep drawing attentions from various fields, ranging from computer science to social science, and to business and industry. With the proliferation of location-aware technologies, geocoded information integrated into social media has become increasingly available. As a typical form of Big Data, social media inherits its high volume, variety, velocity, and veracity (or the Four V's of Big Data

(Gartner, 2018; Schroeck et al., 2012)), which offered unprecedented opportunities for perceiving social dynamics in today's blend of physical and cyber world, yet posed new challenges to the field of geosocial studies.

This chapter begins by introducing the background knowledge of this dissertation (Section 1.1). Then it reviews existing studies on information diffusion regarding its applications, the embedded geographical information, and drives of this process (Section 1.2). Based on the review, research problems on this topic are stated in Section 1.3 and research questions and objectives are raised in Section 1.4. Next, the scope and organization of this dissertation are presented in Section 1.5. In closing this chapter, a framework is designed for answering the research questions and fulfilling the research objectives (Section 1.6).

1.1 Background

1.1.1 Social Media and Twitter

The early Web generally refers to a set of static websites connected by hyperlinks, providing non-interactive contents (Crooks et al., 2014). Then as the Internet became more interactive, Web 2.0 (O'Reilly, 2007) gradually came to the fore. With Web 2.0 technology, users are able to view the web contents like in the early Web era; at the same time they are allowed to contribute information to the sites and to communicate with each other. In Goodchild (2007), Web 2.0 was described as a bi-directional collaboration between the internet contents and the users, with a highlight on user-generated content.

In response to the emphasis of users' participation and user-generated content in Web 2.0, social media was born. Social media is a collective of internet-based applications

built on Web 2.0 technology, and it allows the creation and exchange of user generated content (Kaplan and Haenlein, 2010). It has changed the way news is generated and spread, and has given rise to a better connected global society (Lerman, 2007). Prior to social media age, users communicate by mail, email, message boards, and telecommunication; then this process was simplified to a click of a button. Besides cultivating a new social behavior of its users, social media also prompted the traditional news media to progress by enabling them to get eyewitness information and to connect with broader audiences.

A framework presented in Kietzmann et al. (2011) defines social media via seven functional building blocks: *identity*, *conversations*, *sharing*, *presence*, *relationships*, *reputation*, and *groups*. Within the scope of this framework, different social media platforms focus on a subset of these blocks. Twitter, for instance, focuses on short messages sharing (Figure 1.1), since users receive, generate, and share information in Twitter. Sometimes these messages are users' real-time status updates; sometimes they are about broadcasting news received from other users. Thus, Twitter is primarily about *conversation* and *sharing*. And with the growth of mobile phone usage, location-aware information posted on Twitter largely strengthened the functional building block of *presence*.

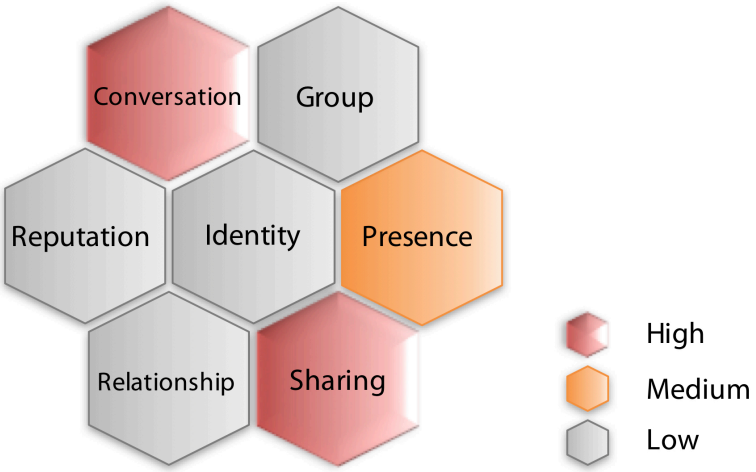

Figure 1.1. Twitter through the lens of the social media building blocks adopted from Kietzmann et al. (2011).

Twitter, as a social media platform, has become a popular venue for online communication and news sharing in the past decade (Suh et al., 2010). In February 2019, Twitter claimed 321 million monthly users (Shaban, 2019). Cox (2016) suggests that Twitter presents far more news items to users than Facebook. In general, users use social media platforms for different purposes, and hence they may behave in different ways on these platforms. One of the greatest differences is in the use of each platform for news broadcasting. Twitter, for example, is unique due to a much stronger emphasis on real-time information. Nearly six-in-ten Twitter news users (59%) use the site to keep up with a news event as it is happening, which is almost double the rate among Facebook news users (Mitchell et al. 2015). Another difference is in the mix of news topics seen on each platform. Twitter users overall see a greater mix of topics, compared to Facebook (Shearer, 2015).

Social media has grown into a promising data source for various purposes such as research, policy-making, and commercial use. But widely acknowledged, that all social media data potentially have biases in who are represented (Malik et al., 2015). Furthermore, despite the large volume of social media data, only a small percentage is geotagged: 2% for Twitter and 25% on Instagram for example (Flatow et al., 2015). The capability of representing the population at large has been doubted when learning knowledge and drawing conclusions from these geotagged data. However, the situation of lacking geotagged information in social media did not stop researchers from making use of this type of data. While geotags offer the most precise location information, we can use other salient information (e.g., hashtags, check-ins, and user's profile) in a social media record to infer its location. Various efforts have been made for this purpose. For example, Palpanas and Paraskevopoulos (2015) analyzed the content similarities of non-geotagged tweets and geotagged tweets, and figured out where the tweets without geotags are posted. In addition to social media content, user's timeline information is also useful for identifying location information (Li et al., 2018a).

1.1.2 Information Diffusion

Information diffusion is defined as "the process by which a piece of information (knowledge) spreads and reaches individuals through interactions" (Zafarani et al., 2014). The term diffusion refers to a universal process of social change, and the diffusion of information is ubiquitous in our everyday life. Therefore, studying information diffusion is an effective way for understanding the dynamics of human communication and interactions among them. For this purpose, studying the mechanism and drives of

information diffusion is essential, for instance how quickly information spreads, how effective the diffusion has influenced the population, and the drives and causes of the diffusion process.

Normally, an information diffusion process is characterized in two dimensions: its structure and its temporal development. Sometimes a third dimension—spatial dynamics—arises when geographic change is considered. Specifically, its structure refers to the environment (e.g., network or a physical location) where information diffusion takes place, which may be static or dynamic. Temporal development captures the evolution of the diffusion over time. Spatial dynamics are normally accounted for by locational change of the information. A large family of existing work on Twitter primarily focus on the first two dimensions: the cyber space where information diffusion occurs, and changes caused by users' status (activate or inactive) as well as the topology of the user networks (Anderson et al., 2015; Hale, 2014; Kim et al., 2018), but the third dimension has not been as widely and thoroughly explored. This is where this dissertation dedicates to.

1.1.3 Space and Place

Concepts as space and place lie at the core of geographical discipline (Tuan, 2001). Adams (2011) stated that space is a rather abstract idea, which "evokes abstraction, inhumanness, meaninglessness, and emptiness;" on the contrary, place is often deemed as the essence of meaning, experience, stability, and coherence. Space offers place position and orientation, while place gives space character and structure.

In the 1970s, geographer Edward Relph addressed that the employment of various media and communication technologies, have encouraged the emergence of 'anti-spatial'

and ‘placelessness,’ indicating a weakening nexus between human individuals and their activities and the social spaces and places they are located in. The geographical constraint was enfeebled since the development of mobility and technology within media has freed people from face-to-face contact when it comes to information transmission: all information is accessible everywhere and anywhere (Relph, 1876). Resonance of such point of view was found in succeeding work (e.g., Graham & Marvin, 1995; Mitchell, 1996). However, some scientists think otherwise. For example, Soja (1985) addressed the important role space plays by stating that “social life is materially constituted in its spatiality.” Harvey (1984) insisted that the introduction of concepts of space into any social theory should have powerful influence on that theory’s core propositions.

Debates about the role space, place, and distance in communications have been triggered and carried on since decades ago. As suggested by Ek (2006), concepts like ‘placelessness’ should be handled with care. Graham (1998) thoroughly explained the conceptual foundation and theory of the roles space and place played in information technologies from both sides of the debate. And evidently, he held the viewpoint opposite to what was stated by Relph (Relph, 1876). He suggested the concept of co-evolution: instead of the substitution and transcendence perspectives of technology towards space and place, new technologies co-evolved with the production of space, place, and human territory. Furthermore, Graham addressed that novel telecommunication forms actually represented and articulated real space and place, encouraging and generating physical mobility, contact, and interaction in today’s highly mobile social world. The social

networking app Meetup[1], for example, successfully integrated the online groups formation and offline meetings within a particular metropolitan area. The fast-growing social activity flash mob is another example of organizing via telecommunications and social media, and grounding the activities by a group of people in real world (CNN, 2009).

1.1.4 The Roles of Geography in Information Diffusion

A widely accepted definition of place is spatial location that has its meaning via human experience (Tuan, 2001). It is quite intuitive that traditionally, humans experience refers to their activities in physical space, and these activities establish meaningful places with unique themes. Nowadays the advancements of new technologies have made humans experience manifold, which characterize geographical information in novel ways. Jenkins et al. (2016) proposed that the emergence of user-generated content such as social media largely contributed to the process of forming a place and shaping its characteristics, with users' collective sense of place. They also discovered the spatial alignment between social media hotspots and corresponding physical locations. Similar work was done by Mok et al. (2010). The authors systematically explored the role of distance in social networks pre- and post-Internet, and found different sensitivity levels of personal relationships to distance. Specifically, email communication was generally insensitive to distance, but tended to decrease slowly over distance; while frequencies of face-to-face and phone call contact dropped significantly over distance. These results support the statement that geography still matters, and as important, in the age of the Internet.

[1] https://www.meetup.com/

In geography, Tobler's First Law (Tobler, 1970) addresses the effect of physical distance, while in societal studies homophily describes similar phenomenon in social space (McPherson et al., 2001). Geographic and hemophilic similarities are essentially connected, as one of the greatest drives of homophily is physical proximity (Hannigan et al., 2013). Based on the experiments conducted in their spatial social network community detection research, Hannigan et al. (2013) introduced an axiom that "as the geographic space of interaction for a social network shrinks, it is more likely that those left within the community are more connected." In other words, the expansion of space weakens the connections between people located in it. This proposition agrees with that was suggested in Wellman (2001), where human interactions at different levels of geographical scales were thoroughly discussed. In addition, homophily can also be shaped by geographical characteristics other than physical distance; for example people residing in similar climate would likely adopt similar living habits, behavior, and fondness (Falconer, 1781).

A large family of work on online social networks have highlighted the cyber environment, where users' communities are formed and activities such as news spreading take place (Hale, 2014; Kim et al., 2018). Meanwhile, the growing awareness of the value of geo-references in social media has inspired the emergence of research stressing the geographical environment, since a variety of users activities are deeply embedded in it (Ferrara et al., 2013; Java et al., 2007; Kulshrestha et al., 2012; Pruthi et al., 2015). Therefore, geography is important for understanding information diffusion. While this type of information used to be unavailable at large scale; nowadays, it is much more easily accessible owning to the widely use of electronic devices. According to comScore (2017),

81% of all social media time is spent on mobile devices in 2016, while five years ago in 2011, it was only half of the percentage (42%). Social media services are extensively adopted on mobile devices, providing rich contents with geographical information.

1.2 Related Work

1.2.1 Information Diffusion

Information diffusion is a vast research domain and has attracted interests from various disciplines, such as physics (Zhang et al., 2016), biology (Chen et al., 2018a), business (Agarwal et al., 2019), and public policy (Zhu et al., 2018). Among them its applications in social sciences is of my interest. Questions raised in societal studies are usually as follows: (1) how and why information is diffusing now and in the future; (2) what kind of information is popular and diffuse the most; and (3) which participants in the population play important roles in the diffusion process.

Organized thematically, applications in group (1) include: inferring and predicting the structure of information cascades (e.g., Galuba et al., 2010; Molaei et al., 2019), information diffusion models selection and evaluation (e.g., Saito et al. 2010; Zhang et al. 2013), and community detection (e.g., Adamic and Glance, 2005; Ramezani et al., 2018); applications in group (2) include: topic analysis and trending topic detection (e.g., Chae et al. 2012; Ferrara et al. 2013), maximizing the spread of information/influence (e.g., Gomez-Rodriguez et al., 2012; Wang et al., 2018; Yerasani et al., 2019), and minimizing the spread of misinformation (e.g., Budak et al., 2011; Tan et al., 2019); and applications in group (3) include: seed selection for further diffusion (e.g., Kim et al., 2014; Li et al., 2018), and social media marketing (e.g., Woo and Chen 2016; Zhao and Li 2019).

1.2.2 Geographical Information in Twitter

The geographical characteristics of Twitter networks have been a major interest to researchers. The geographical distribution in the Twittersphere was investigated in Java et al. (2007). Agarwal et al. (2018) investigated the geographical distribution of sentiment in Twitter. Kulshrestha et al. (2012) stressed the substantial impact of geography on user interactions in the Twitter social network. To discover the information diffusion process, De Choudhury et al. (2010) examined several data sampling methods on a set of node attributes including location. Results revealed that a sample that incorporated users' location could improve their model by a significant margin. Ferrara et al. (2013) characterized the relationship between trends and geography via a network depicting the conversations flow on Twitter. They identified two main classes of trending topics, both in unique patterns geographically: those that surface locally, coinciding with three different geographic clusters; and those that emerge globally from several metropolitan areas. The emergence of these geography-characterized classes suggests that the nature of information diffusion through Twitter is deeply embedded in geographical locations.

Among all geographical characteristics, geographical scale has drawn wide attention. Hannigan et al. (2013) illuminated that the expansion of space weakens the connections among people located in it. This proposition agrees with that was suggested in Wellman (2001), where human interactions at different levels of geographical scales were comprehensively discussed. Also, Takhteyev et al. (2012) concluded that distance matters in Twitter activities, at both short and long ranges. Pruthi et al. (2015) discovered the different influence of distance between an event's location and Twitter users' location in

regional and global events. Another important geographical attribute is location, because events are strongly localized at place, and so are the participants involved in them. As explained earlier, novel online communication forms articulated geolocation (Graham, 1998); and it is endorsed by a more recent work, where the user-generated cyber-social events are found useful for representing the urban landscape in physical terms (Crooks et al., 2016). In addition, Jenkins et al. (2016) discovered the spatial alignment between social media hotspots and corresponding physical locations. Therefore, it is advised that a variety of geographical attributes need to be inspected when studying Twitter activities with geographical descriptions.

1.2.3 Drives of Information Diffusion

Understanding what stimulates the diffusion of information is meaningful, especially for the purpose of targeting specific groups of people, controlling diffusion directions, and expanding/shrinking diffusion scales. In existing research, information diffusion caused by social influence and homophily has been widely studied considering their combined and respective effects.

Social influence, defined as a social phenomenon that individuals can exert and receive, and that induces similar behaviors or decisions to their connections, is considered a significant drive of information diffusion (Guille et al., 2013). When social influence becomes high, some pieces of information might become extremely popular, spread to longer distance in a short time, and become more influential and generate new trends. Homophily, the tendency of individuals to connect to similar ones, is another drive of

information diffusion (McPherson et al., 2001). Such similarity can be age, gender, interest, and location.

Bakshy et al. (2011) suggested that despite the wide availability of data, identifying influence remains a challenge. Observational data shows that individuals tend to engage in similar activities as their peers; however it is often impossible to determine whether a correlation between two individuals' behaviors exists because they are indeed similar or because one person's behavior has influenced the other (La Fond and Neville, 2010; Simons-Morton and Farhat, 2010). Aral et al. (2009) aimed to distinguish influence- and homophily-driven behavior adoption in dynamic networks. They found that previous studies tended to overestimate peer influence, and that homophily explained more than 50% of the behavior contagion. Mislove et al. (2010) investigated users' attributes in an online social network, and concluded that homophily played an important role in community formation. Hannigan et al. (2013) claimed that geographic and hemophilic similarities are essentially connected, as one of the greatest drives of homophily is physical proximity. Therefore, when exploring the role of geographical characteristics in information diffusion, it is beneficial to distinguish them from other influential factors. However, it is challenging to separate them since social relations are deeply embedded in the physical world.

1.3 Problems Statement

As is clear from the above review, information diffusion in online social networks has been extensively researched for varied purposes; yet challenges and research gaps still exist.

This section focuses on stressing the difficulties and gaps of the interested study area, and further leads to the research questions and objectives of this dissertation (Section 1.4).

1.3.1 Complexities Added by Geographical Information

In online social networks, information diffusion happens through cyber space interactions among users, and also has a physical space presence as users have to be somewhere. When approaching the spreading process solely in networks or in physical space, each step can be easily understood and represented: a series of hops in a network or locational moves in physical space. However, when interpreting the information diffusion process in the nested cyber and physical dimensions, undoubtedly complexities will be added. As social media contents are becoming increasingly geo-located, additional context (i.e., physical environment) is presented, as well as the emerging demands for understating the corresponding analysis and processing, such as locations and their variations over time (Croitoru et al., 2014).

The first complexity is the expression form of geographical information: coordinates, distance, or descriptive toponym such as city/country names. Besides the selection of proper spatial expressions, these forms of spatial information require different methods for incorporating geographical information into the analytical method; for example as a variable, a parameter, or a constraint. This difficulty is mostly due to the complex nature of social media data. Second is the practical meaning of the results generated from the added spatial information: if the geographically specified or constrained generalizations have any realistic meaning and significance, and how to explain such meaning and significance; for example aggregated users behavior caused by ground-based

events. Third is the distance and scale issue in geographical space: how to interpret information diffusion at varied geographical distance and scales distinctively is inevitably a challenge. Fourth is the interweaving of geographical influence and other influential factors on information diffusion. In the context of this research, it is important to recognize the distinctions and relations between geographical influence and other influential factors. However, it is challenging to isolate the geographically associated factors, since the networks are deeply embedded in the physical world.

1.3.2 Gaps of Studies on Georeferenced Information Diffusion

Though geographical influence on Twitter activities is already well-acknowledged, further exploration of geographical influence on information diffusion is still needed. Previous studies on information diffusion considering geography mainly focused on the changes of reached location and coverage area (Kwon et al., 2015; Pruthi et al., 2015; Puri et al., 2018). However, interactions within the Twitter users at different locations at a specific time is unknown. This issue is identified as information flow. In limited number of studies involving information flow in Twitter, it is usually utilized for visualization (Croitoru et al., 2015; Lotan, 2011; Mishori et al., 2014); while mining meaningful patterns in it is often overlooked, especially in the geographical space. Understanding information flow pattern in Twitter is important because it depicts the internal mechanism of information diffusion, surfacing the essential patterns veiled in the mass data, and facilitating decisions and further applications in an accurate, responsive, and flexible manner.

1.3.3 Big Data Challenges

The last challenge lies in the rapid expansion of information we access today. Different from the time when word-of-mouth and face-to-face communications are the major ways for information exchange, nowadays we live in the Information Era and are exposed to vast new information daily. Big Data, as presented through social media, has generated opportunities to perceive social dynamics in novel ways. Mining key values in Big Data, such as the new-found challenges in finding and using the hybrid mix of spatial and social contents in social media (Croitoru et al., 2017), is commonly referred to as the needle-in-the-haystack problem (Kaisler et al., 2013), indicating the difficulty of achieving the desired results.

The Big Data challenge has drawn substantial attentions in various fields such as computer science, social science, IT industry, and the analytics industry. As a typical form of Big Data, social media inherits its high volume, variety, velocity, and veracity (or the Four V's of Big Data (Gartner, 2018; Schroeck et al., 2012)); and commonly provides data in the form of streams. These characteristics of social media have offered unprecedented opportunities for perceiving social dynamics in today's blend of physical and cyber world, yet posed new challenges to the field of social studies. First, the increasing amount of data has raised an immediate challenge to traditional analytical environments and methods. Second, various forms of social media such as images, texts and videos, require extra work to unify the unstructured data for downstream usage (Croitoru et al., 2013). Third, velocity, the speed at which data is produced and processed, has urged the invention of new algorithms and methods to properly handle streaming data produced from social media

platforms (Bello-Orgaz et al., 2016). And fourth, veracity emphasizes the inherent uncertainty of social media data (Schroeck et al., 2012), and the capability of distinguishing useful pieces from misinformation is crucial (Tan et al., 2019).

Narrowing down the four V's of Big Data to the context of this dissertation, major challenges exist in (1) large volume—data organization and management, and information visualization, (2) high variety—standardizing different representation forms of variables, (3) high velocity—evolving social media data streams, and (4) high veracity—noise cleaning and irrelevant and false information removal. In response to these challenges, a framework relieving these difficulties is needed.

1.4 Research Questions and Objectives

With the growing popularity of online social media services such as Twitter and Facebook, it becomes more and more important to understand how users communicate on these platforms. Targeting this issue, this dissertation focuses on mining the spatiotemporal patterns of information diffusion in online social networks. More specifically, it proposes the following research questions:

- What are the spatial distribution and temporal change of information diffusion in online social networks?
- Does geography matter in an information diffusion process in online social networks?
- How to interpret the discovered patterns and transform our observations to real-world knowledge for informed decision-making?

Solving these intriguing problems is challenging yet will facilitate our understanding of the diffusion process of information, regarding its structure and organization, its spatial pattern, and its evolvement over time. To find solutions for these research questions, this dissertation will firstly develop a suitable method for organizing the available social media data and obtaining pre-processed products required for the ensuing spatiotemporal analysis; second, explore the spatiotemporal patterns of information diffusion formed by communications in online social networks, and third, seek the social and geographical drives of the information diffusion process. The first objective is fulfilled by developing a density-based stream clustering method based on a previously established algorithm (Chapter 2). The second objective is approached from the structural, geographical, and temporal analysis of the information diffusion process in the nested network topology and geographical space (Chapter 3). The third objective is accomplished through further investigations adding real-world facts and perceptions (Chapter 3). Successful analysis will augment our understanding of information diffusion in cyber and physical spaces.

1.5 Dissertation Scope and Organization

This dissertation presents two lines of research topics: one focuses on streaming clustering and the other concerns big data mining. In the first research, an adaptive stream clustering method for large data streams with location information is developed. It has not only filled the gap of consistently acquiring real-time point-cluster relations during the clustering process, but also improved existing algorithm implementations by alleviating memory constraint, resolving data point overlap, and recovering false noise. All these advancements

contribute to various application domains with demands alike. In the second research, built on the stream clustering results, spatiotemporal patterns of information diffusion in social networks formed by online communication are discovered, regarding its network structure, volume, direction, location, and temporal evolvement. The second research novels in its core mechanism of information source-to-sink flow analysis based on multi-temporal similarity measure, and contributes to how cyber population react to public health issue via case studies on epidemics.

Discovering the spatiotemporal pattern of information diffusion in online social networks contributes to social system studies, communication, economic and human geography, and public relations and marketing. Sociologists, economists, and social media marketers might find interests in this dissertation. This research also introduces methods for big data organization and clustering, data mining, network analysis, and spatiotemporal analysis. Scientists in the fields of data mining, network, and spatial analysis may also find useful information. Moreover, since the chosen case studies pertain to emerging infectious diseases of international concern, this work offers insights into how online discussion develops on this topic and how cyber users perceive and propagates such information. Therefore, international scientific community and policymakers in public health, international relations, and political science may find this work useful for guiding better policy recommendations and preparations for future epidemics.

This dissertation is composed of four chapters. Chapter 1 introduces the background (Section 1.1), reviews related work (Section 1.2), explains problems and difficulties in existing work (Section 1.3), and states the research questions and objectives

of this dissertation (Section 1.4). The last section (Section 1.6) in Chapter 1 describes the methodology of this dissertation research, summarizing the next two chapters where the two main components of this dissertation are presented: Chapter 2 presents the first research on stream clustering method; and in Chapter 3, spatiotemporal analysis of information diffusion is performed. Chapter 4 provides a conclusion summarizing both studies with suggested future work.

1.6 Overall Framework

Taking advantage of the richness and availability of online social networks, I construct the framework for fulfilling the research objectives based on this type of data, and specifically Twitter is selected as the data source. Next the core subject in this research, "information" needs to be defined. Given the wide-ranging topics in Twitter and their high dynamics over time, it is difficult to include them exhaustively. Thus, a filtration of topics in the information is necessary. Among all topics in Twitter I chose to study public health related topics. Additionally, since geographical aspects in the information diffusion process are major interests, information associated with locations is desired. Under this requirement, event-associated information fits well because usually they are location explicit. Once data type is chosen, a series of analysis can be performed. An overall framework summarizing the methodology adopted in this dissertation is shown in Figure 1.2.

To obtain useful Twitter data, first raw tweet records were harvested through the Twitter API, and organized in a MongoDB database (https://www.mongodb.com/). Then applying a series of filtering criteria and preprocessing steps, I obtained a set of useful tweets. In this study, aggregating individual records into meaningful clusters is essential

because of the large amount of social media data records, and analysis at the aggregated level can reveal the most important pattern while excluding unnecessary information. Thus, on the dataset of useful tweets, clusters were generated from the developed clustering method—GeoDenStrem—using the locations of individual records, and then shaped information flows among their containing individual records. GeoDenStream is the basis for obtaining meaningful clusters using the given Twitter data, and for supporting the subsequent spatiotemporal analysis of information diffusion. Its development process is deliberated in Chapter 2 through its conceptual design (Section 2.3), implementation (Section 2.4), verification (Section 2.5), and application (Section 2.6).

Another important method of this dissertation is the spatiotemporal analysis built on the stream clustering results in Chapter 3. It is approached from two levels: one highlights the source and sink clusters, and the other emphasizes information flow. Here I define the clusters with higher outgoing flows than incoming flows as information source, otherwise as information sink. The source-to-sink information flow provides quantitative and geographical descriptions of information diffusion (Section 3.4). Further discussion is performed upon the analyses of information source and sink, as well as information flow from the structural (Section 3.5.2), spatial (Section 3.5.3), and temporal aspects (Section 3.5.4). In addition, spatiotemporal scales are discussed based on the key parameters setting in the GeoDenStream clustering process (Section 3.5.1).

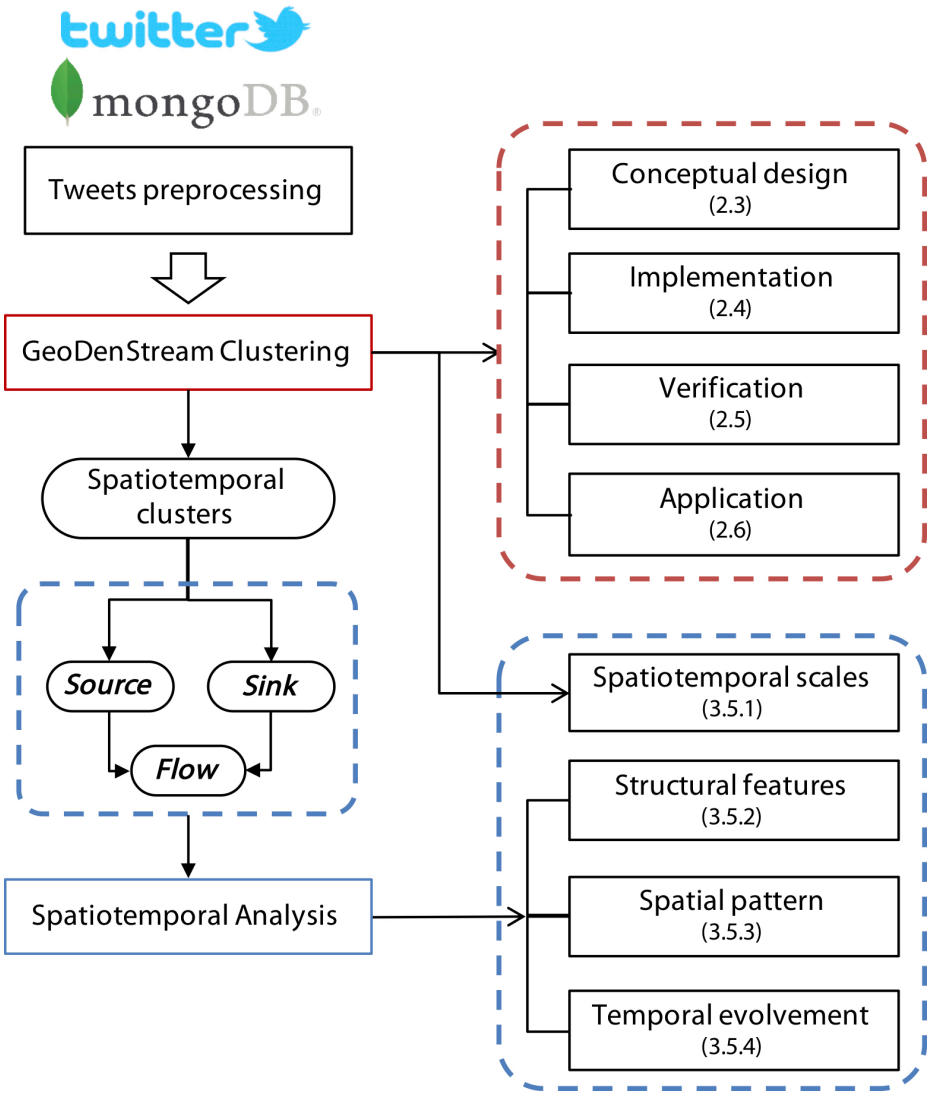

Figure 1.2. Overall framework of this dissertation; section numbers are included to direct where each topic is described.

2 GEODENSTREAM: AN IMPROVED DENSTREAM CLUSTERING METHOD FOR ACQUIRING INDIVIDUAL DATA POINT INFORMATION WITHIN GEOGRAPHICAL DATA STREAMS

Abstract

Data streams with location information are prevalent nowadays due to their close alignment with real-world events and phenomena. To satisfy the demands for organizing and analyzing individual data records within such data streams, clustering has been widely accepted as an effective and efficient tool. Existing implementations of DenStream, a popular density-based stream clustering method, have major drawbacks that the point-cluster relationship is untraceable, and individual point information in the clusters is not recorded. To fix these weaknesses, this chapter proposes an improved DenStream method, named GeoDenStream, aiming at finding cluster shapes, maintaining information of individual points in the clusters, supporting spatiotemporal analysis based on the explicit point-cluster relationships, and facilitating the discovery of patterns hidden in the data. Specifically, the design of GeoDenStream is elaborated, its performance is verified by two synthetic datasets, and its practicability is tested on two case studies, where spatiotemporal analysis is accomplished using the clustering results. Results show that the inherent difficulties in implementing DenStream are greatly alleviated by GeoDenStream.

2.1 Introduction

In recent years, data streams have become an integral part of the rapidly evolving modern information landscape. Various application domains, such as health (Althouse et al., 2015), transportation (Liu et al., 2011), finance (Liu et al., 2010), communication (Naaman et al., 2010), energy (Vikhorev et al., 2013), climate and weather (Freeman et al., 2017), and environmental monitoring (Funk et al., 2015), produce real-time data streams, and rely heavily on the availability of near-continuous data flows for higher level reasoning and decision making (Valle et al., 2009). In many of these domains, data streams are closely associated with human activity in geographical spaces. Examples of such activity-driven streams range from a user's (entity) check-ins and check-outs at access-controlled facilities (Kromwijk et al., 2010), to users' GPS-enabled movement tracking streams (Moreira-Matias et al., 2016) and geotagged content sharing in social media (Stefanidis et al., 2013). The tight coupling between space, time, and activity in such streams can potentially provide a rich source of information about human behavior and activity patterns. This potential and emerging need to analyze such streams, which has fostered a growing interest within the data mining community (Atluri et al., 2017), serves as the primary motivation for the work presented here.

Generally, it is possible to conceptualize a data stream S as consisting of a sequence of n ($n \rightarrow \infty$) time-stamped records $(X_1, t_1), (X_2, t_2), \ldots, (X_n, t_n)$, where each record X_i is comprised of a set of d attributes $\{x_i^1, x_i^2, \ldots, x_i^d\}$, and t_i is a time stamp indicating when the record was created or received (Aggarwal et al., 2003). While the record attribute vector can contain any type of attribute information, this chapter explores the analysis of streams

in which at least one of the record attributes contains geographical information (e.g. geographical coordinates or a toponym). In the remainder of this chapter I use the term geographical data stream to denote such a stream. Notably, geographical data streams are spatiotemporal in nature as they combine spatial and temporal information in a single stream record. Additionally, it is important to note that a data stream can, in general, be dedicated to capturing data about one of two types of constructs: entities and events (Krempl et al., 2014). Here, the term entity relates to a discrete thing that endures over time, e.g. a building, a vehicle, or a person, while the term event relates to an occurrence in space and time, e.g. the detection of smoke at a particular sensor location or the detection of congestion along a highway. In practice, a key difference between entity data streams and event data streams is that the former must include a unique entity identifier (e.g., a vehicle ID), while the latter may not. When dealing with entities it is also important to recognize that entity stream data can be analyzed at different levels of granularity, from the discrete entity (a person moving in geographical space) level to groups of entities (e.g., a group of people moving together).

Given a geographical data stream, it is often of interest to analyze them in order to derive higher level information that would support reasoning and decision making. Such analysis can include a wide range of operations, from basic data analytics, to clustering, pattern and entity mining, event detection, and process modelling (Krempl et al., 2014). Among these, clustering has emerged as one of the most commonly used analysis operations (von Luxburg, 2007; Xu and Tian, 2015). As a result, various stream data clustering algorithms have been proposed based on a range of data models and similarity

(or distance) measures (Gaber et al., 2005), which can be broadly organized into 5 primary classes, namely Growing Neural Gas (GNG) methods, hierarchical methods, partitioning methods, density-based methods, and grid-based methods (Ghesmoune et al., 2016).

Selecting an algorithm from one of these classes is not always straightforward due to the underlying difficulty in defining a universal notion of a cluster that can be applied in any context. Furthermore, the algorithms in each class may rely on a different set of assumptions, criteria, and underlying model. Consequently, the selection of the clustering process often tends to be domain specific and exploratory in nature (Estivill-Castro, 2002). When clustering geographical data (and data streams) density- or grid-based methods, such as DenStream (Cao et al., 2006), StreamOptics (Tasoulis et al., 2007), or FlockStream (Forestiero et al., 2009), are often selected (Xu and Tian, 2015). This selection can be attributed, at least in part, to two primary reasons. First, the concept of density naturally lend itself to spatial and spatiotemporal domain since in these domains the notion of a cluster is often associated with the "high concentration" of data points. Second, density-based clustering methods offer several distinct characteristics that are advantageous when dealing with activity-based data. Specifically, density-based methods (i) do not require a priori information about the number of clusters, (ii) can handle clusters with arbitrary shapes, and (iii) detect and handle outliers (Amini et al., 2014).

Another important issue that should be addressed when selecting a clustering method is the way in which cluster information is maintained and reported. In some clustering methods the focus of the process is to detect whether one or more clusters exist, and when clusters are detected to report and preserve only key summary descriptors about

each cluster. An example of this approach can be found in the algorithm presented by O'Callaghan et al. (2002), in which only the centers of clusters are maintained over time as stream data is processed. Similarly, in the framework presented by Aggarwal et al. (2003) only information about the center and radius of each micro cluster, along with unique cluster IDs are maintained over time. A key advantage of maintaining only summary descriptors is that it enables managing the clustering process efficiently since each cluster, which can potentially include a large number of stream records, is described only by a limited set of data-driven parameters (e.g. center coordinates and a radius). Such an approach, however, is not suitable for streams that observe discrete entities over time, such as moving vehicles, travelling individuals, or the geotagged postings of a social media user, since the clustering process does not maintain the entity-level composition of each cluster over time. The challenge I address in this chapter is therefore how to adapt the commonly used offline-online phase density-based clustering to support entity stream mining.

In view of these considerations, this chapter proposes a method for enhancing existing density-based stream clustering methods in order to support entity stream mining in geographical space. For this purpose I build on DenStream, a density-based clustering method presented by Cao et al. (2006). The selection of DenStream in this chapter is based on three key considerations. The first consideration relates to the conceptual framework behind it: DenStream is based on the conceptual framework for clustering evolving data streams proposed by Aggarwal et al (2003), which involves the creation of micro and macro clusters in a two-step (online and offline) processing approach that is employed in many stream clustering methods. The second consideration relates to the historical development

of density-based stream clustering methods: Since its introduction in 2006, DenStream has served as the foundation for the development of various other density-based algorithms. A third consideration relates to the availability of the DenStream Algorithm. Consequently, I argue that because of these considerations the enhancements of DenStream I propose could, in principle, be more easily adapted to enhance other density-based stream clustering methods to support traceable spatiotemporal clustering.

The remainder of this chapter is organized as follows. Section 2.2 provides an overview of the DenStream algorithm and describes in more detail its key limitations in the context of entity stream mining in geographical space. Building on DenStream, its limitations, and the considerations noted above, Section 2.3 introduces GeoDenStream, a density-based stream clustering algorithm that supports entity stream mining in geographical space. Its implementation is stated in Section 2.4. In 2.5, two synthetic datasets were tested on GeoDenStream to verify its performance. To showcase the utility of GeoDenStream, Section 2.6 summarizes the clustering analysis of two real-world Twitter datasets. Finally, a discussion and summary of these results is provided in Section 2.7.

2.2 An Overview of DenStream

2.2.1 Conceptual Framework

In order to conceptualize the DenStream algorithm in the context of entity stream data in geographical space consider a data stream in which each record is comprised of a data "point", i.e. a geographic location (for example, in the form of geographic coordinates), a time stamp, and a set of related attributes that describe an entity. The DenStream clustering

method applies the core-micro-cluster approach to detect arbitrary-shaped clusters (Cao et al., 2006). In this approach, a core-micro-cluster is constructed by points that are sufficiently dense according to a density threshold, and such cluster evolve over time as data is received. In addition, core-micro-cluster are assigned a weight that decreases exponentially with time. Based on their weights, core-micro-clusters with higher weights (i.e., potential-clusters) are acquired for building clusters, and core-micro-clusters with lower weights (i.e., outlier-clusters) are removed from the final clustering results.

There are four phases in the original DenStream clustering method, as show in Figure 2.1: an Initializing phase in which the potential-cluster and outlier-cluster lists are constructed; an Online phase in which newly arrived data points are either merged into a potential-cluster or form a new outlier-cluster; a Pruning phase in which potential- and outlier-clusters with lower weights are removed from the corresponding list; and finally, an Offline phase in which DBSCAN clustering (Ester et al., 1996) is used for generating offline-clusters based on the potential-cluster list. In this process, a set of parameters are used, including *initial_points* and *min_points* in the Initializing phase; *epsilon*, *lambda*, *beta*, and *mu* in the Online phase; *tp* in Pruning phase; and *offline* in the Offline phase. A more detailed description of these parameters is provided in Table 2.1.

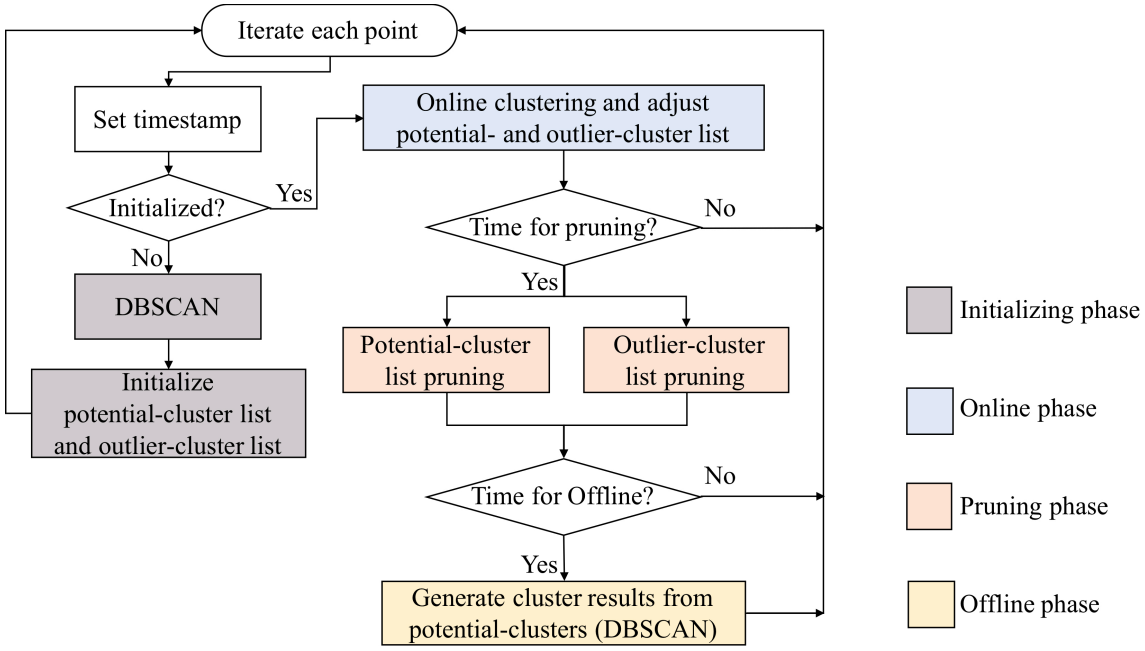

Figure 2.1. Processing flow of DenStream clustering method for geographical data stream.

Table 2.1. Parameters in the original DenStream clustering method.

Phase	Parameter	Description	Value range
Initializing	*initial_points*	Number of points for initializing potential- and outlier-cluster list	1~maximum integer
	min_points	Minimum points for constructing a cluster when initializing with DBSCAN	1~maximum integer
Online	*epsilon*	Maximum radius for a core-micro-cluster	Float number larger than 0
	lambda	Decay factor for weight	0~1
	beta	Weight threshold for outlier-clusters	0~1
	mu	Weight threshold for core-micro-clusters	0~maximum float value
Pruning	*tp*	Time interval for pruning	Any rational time interval
Offline	*offline*	Multiplier of *epsilon* for meaningful clusters in DBSCAN	2~maximum spatial range of the input data stream

2.2.2 Limitations

While the original DenStream algorithm can be used to cluster event streams, it does not explicitly support entity stream mining. The key reason for this limitation is that the algorithm focuses on deriving the location and approximate shape of core-micro-clusters to represent the clustering results rather than keeping track of the relationships between entities and clusters across the clustering iterations. As a result, three major issues need to be considered:

(1) Memory requirement: DenStream can handle stream data with limited memory by applying periodic pruning in which clustering result at each iteration are represented by summary cluster information (e.g., center and radius). However, when it is necessary to

track points (entities) across iterations, limiting memory usage needs to be considered once again. A mechanism for maintaining track of across clustering (and pruning) iterations is therefore needed.

(2) Point data overlap: generally, it is possible that the footprints of some neighboring potential-clusters will overlap during the clustering process. DenStream handles such overlaps when a cluster is pruned by assigning any data points within the overlapping area to a cluster that is not pruned (Cao et al., 2006). However, this strategy does not support entity stream clustering. To illustrate this, consider the scenario depicted in Figure 2.2, in which the evolution of two neighboring potential-clusters, MC_A and MC_B, is shown between time stamp T1 and a later time stamp Tn. In this scenario I assume that the Offline phase is not activated between T1 and Tn. In Figure 2.2 (a) at time T1, an attempt is made to merge the point P1[T1, x, y] to the nearest potential-cluster based on Euclidian distance to the cluster centers, which will result in the merging of P1 to MC_A as shown in Figure 2.2 (b). Let us further assume that over time MC_A and MC_B evolve as new data is received, resulting in the clusters (and their corresponding centers) shown in Figure 2.2 (c) at time Tn. Consider now a newly received data point P2 [Tn, x, y] with the same coordinates with P1. This point will be merged into cluster MC_B instead based on the shortest distance criterion. Therefore, although P1 and P2 share the same location, they belong to different potential-clusters after the Online-Offline cycle. This would result in an ambiguity in entity stream clustering since data points related to the same entity would belong to different clusters at the same time.

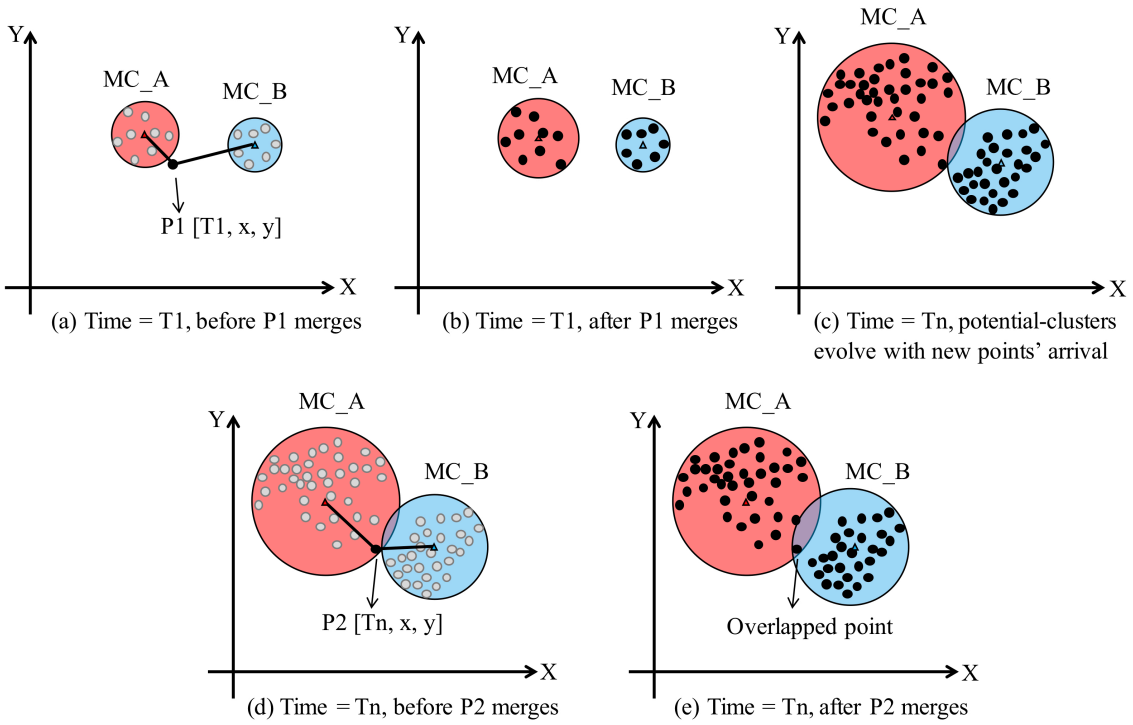

Figure 2.2. Overlap issue in DenStream clustering method.

(3) False noise: periodic pruning operation in DenStream can substantially relieve the memory requirement by classifying low-weighted core-micro-clusters as noise and removing them from the potential- and outlier-cluster lists. However, such pruning may result in false noise points, as shown in Figure 2.3. Specifically, in Figure 2.3 (a), few points are in the vicinity of point P1 at time T1, and so P1 is treated as an outlier-cluster. Then pruning is activated at time T1+*tp*, resulting in the removal P1. Later, at time Tn (Tn >> T1+*tp*, and time lapse between T1 and Tn is still within the time window of an Online-Offline cycle), as points begin to appear around P1 a new potential-cluster is formed. P1 could have been merged into this new potential-cluster since Offline phase has not yet

started, but it was already removed earlier, resulting as a "false noise". Although the removal of such false noise is not likely to affect the shapes of clusters generated in the Offline phase, it may affect the results of analysis within clusters with respect to its entities.

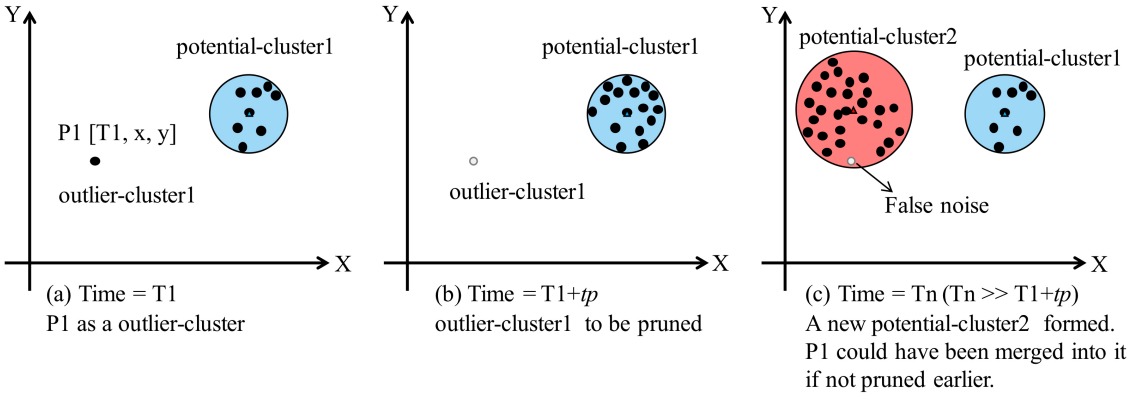

Figure 2.3. False noise caused by periodic pruning.

Despite their significance for entity-based clustering, the limitations noted above have not been explicitly addressed in various enhancements to DenStream that were recently introduced. Such enhancements, as summarized in Table 2.2 focused primarily on improving distance calculation (e.g. HDenStream), parameter selection (e.g., SOStream), and high dimensional data support (e.g., HDDStream, PreDeConStream, and FlockStream).

Table 2.2. Improved DenStream clustering methods.

Method	Improvements	Results
C-DenStream	Adds constraints for certain points	Arbitrary shape of clusters
rDenStream	Adds a retrospect phase for false noise	Arbitrary shape of clusters
SDStream	Extends sliding window for more recent data	Arbitrary shape of clusters
HDenStream	Supports distance calculation of categorical data	Arbitrary shape of clusters
SOStream	Supports automatic calculation of parameters	Clustering parameters and arbitrary shape of clusters
HDDStream	Supports high dimensional data	Arbitrary shape of clusters
PreDeConStream	Supports high dimensional data	Arbitrary shape of clusters
FlockStream	Merges online and offline phase	Arbitrary shape of clusters

C-DenStream mainly focuses on adding user-specified constraints for assigning points to clusters, and its online-offline process aligns with that in the original DenStream (Ruiz et al., 2009). rDenStream adds a retrospect phase to handle false noise, which is similar to the Post-processing phase in GeoDenStream (Liu et al., 2009). Nevertheless, it lacks consideration of memory limitation and points' overlap. SDStream uses a sliding window to process the most recent data and to summarize old data (Ren and Ma, 2009). HDenStream adds a distance calculation method for categorical variables, in order to support categorical and continuous data (Lin and Lin, 2009). SOStream generates parameters required by DenStream based on the Self Organizing Maps method, which has proved to be time-consuming (Isaksson et al., 2012). HDDStream (Ntoutsi et al., 2012) and PreDeConStream (Hassani et al., 2012) extend the basic DenStream to support high

dimensional data. FlockStream employs a bio-inspired model to enhance the efficiency of merging points into clusters, and combines the Online and Offline phases (Forestiero et al., 2009). However, extra work is needed since this method does not offer any noise removal strategy.

2.3 GeoDenStream

Motivated by the limitations of DenStream with respect to entity stream clustering, GeoDenStream focuses mainly on generating clusters while maintaining point entity information across clustering iterations. This is achieved through a novel framework shown in Figure 2.4. GeoDenStream follows a process similar to the original DenStream algorithm, from Initializing phase, to Online phase, Pruning phase, Offline phase, and finally to Post-processing phase. However, these phases are enhanced to better address the memory requirement, data overlap, and false noise limitations noted earlier. These enhancements are described in the following subsections. While in this description an entity is considered as a cluster of geotagged stream records, other record granularity levels can be applied in GeoDenStream using the same approach presented here.

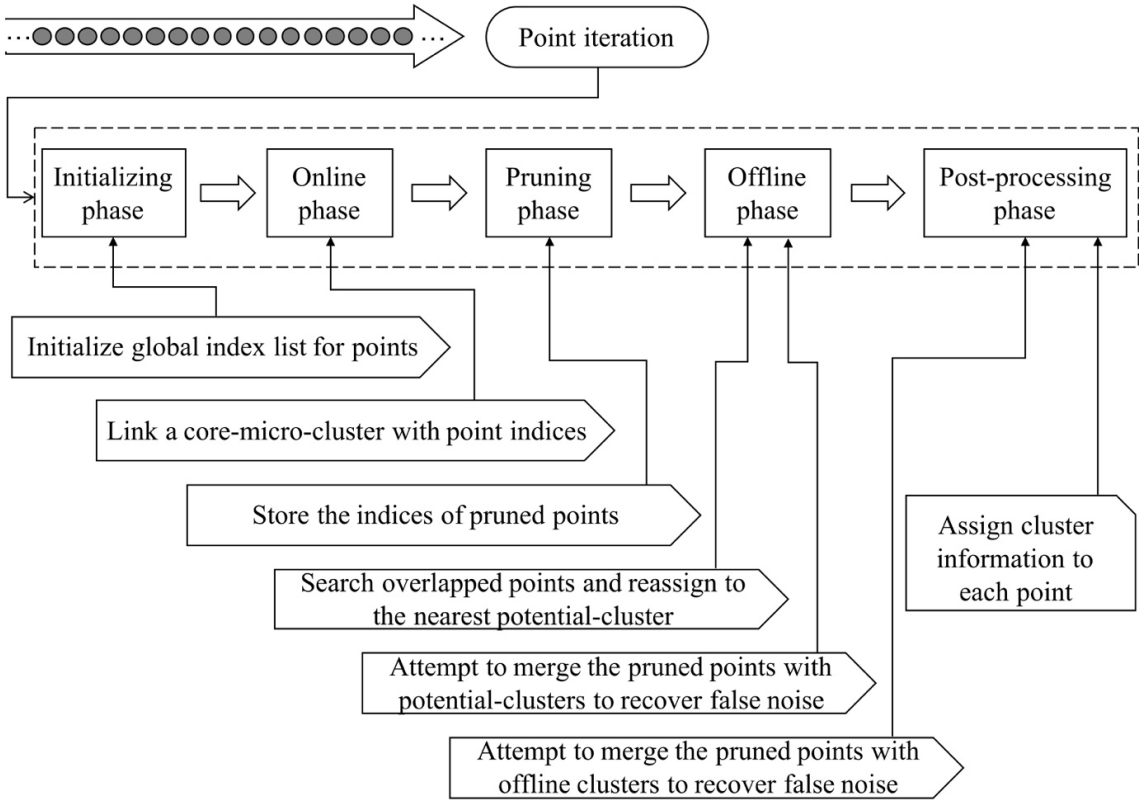

Figure 2.4. Framework of GeoDenStream for analyzing geographical data streams.

2.3.1 Indexing Stream Points

In order to consistently track the index of each point in GeoDenStream, the index of each point is created based on its arrival sequence. This sequence-based indexing method, which is conducted during each clustering cycle, links between each data point and its corresponding data record (global index), between core-micro-clusters and its containing points (clustered-index), and between the pruned noise and the points it refers to (pruned index). In order to reduce the memory required to store the index a long integer data type is used, and index structures are reset every clustering cycle.

As shown in Figure 2.5, a global index list storing all point indices is constructed once the Initializing phase begins. It then passes its records to clustered-index lists and pruned-index lists as the Initializing phase continues. Every potential-cluster and outlier-cluster maintains a clustered-index list which stores the indices of points included in it. Periodically, core-micro-clusters of lower weights are pruned; for such pruned clusters, the indices of the points comprising them are saved in a pruned-index list. Based on the clustered-index list of each core-micro-cluster, cluster information such as cluster ID is assigned to its containing points in the Post-processing phase. The pruned-index list is used for false noise recovery, by checking the pre-recognized noise points contained in this list in the Offline and Post-processing phases.

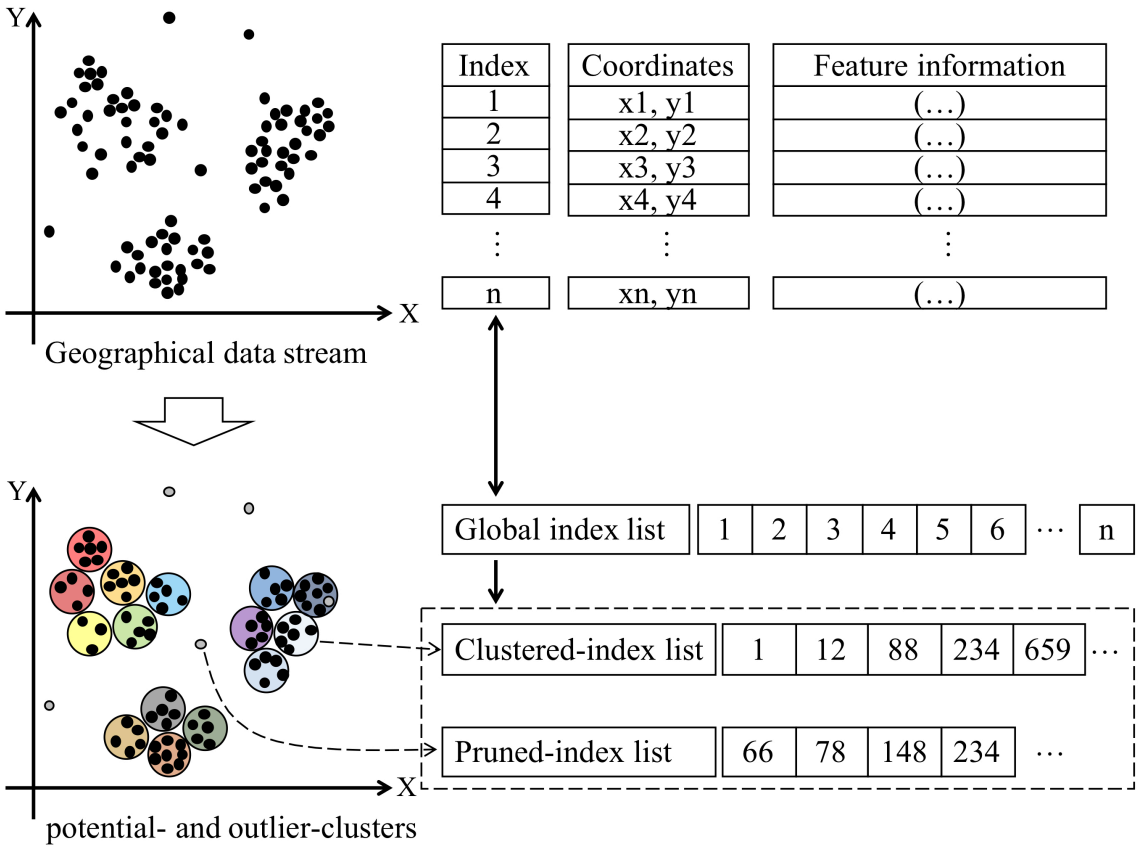

Figure 2.5. Sequence-based indexing strategy.

2.3.2 Overlapping Points Reassignment and False Noise Recovery

In the Offline phase, the generated potential-clusters are re-clustered using DBSCAN. As noted earlier, clusters' evolving process could lead to the situation when points with the same coordinates belong to different potential-clusters. Additionally, as the pruning operation is applied periodically, some pruned points may be falsely labelled as noise. Figure 2.6 depicts the Offline phase with reassignment of overlapping points and recovery of false noise. In order to handle overlapping points, all potential-clusters are checked using

the clustered-index list created in the Initializing phase. If points in different potential-clusters have identical coordinates, their distances to the related potential-clusters will be computed and compared, and then all points with identical coordinates will be reassigned to the nearest potential-cluster.

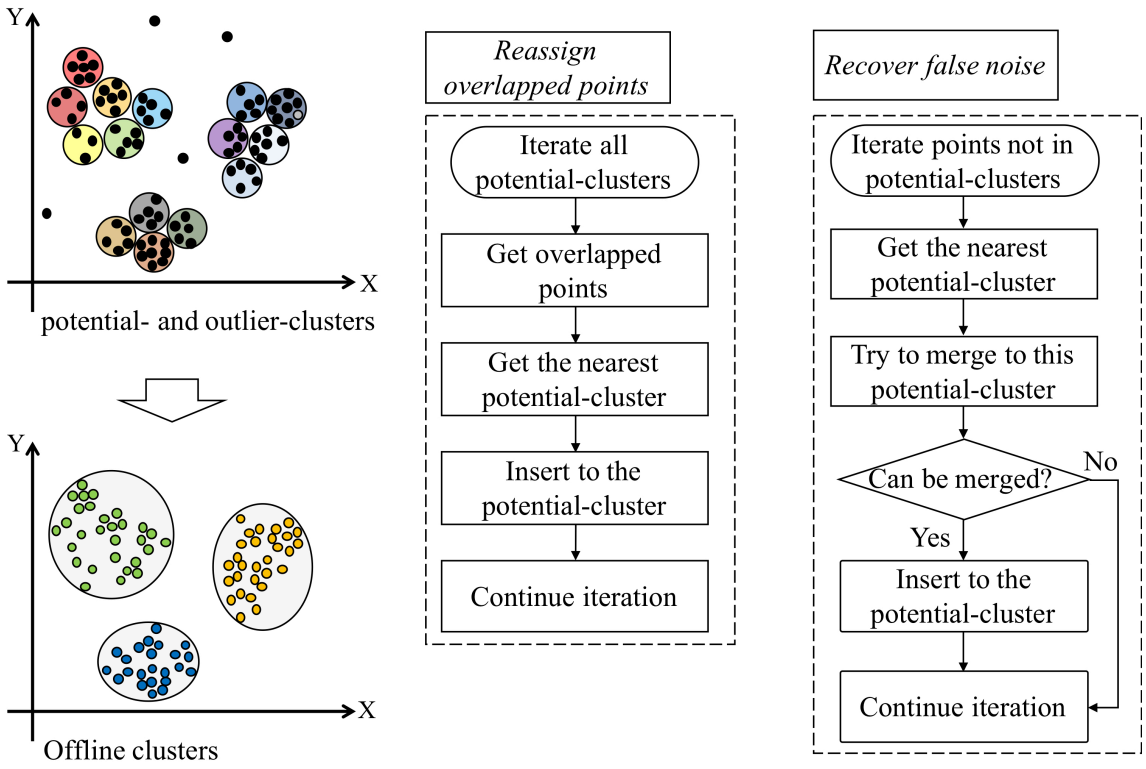

Figure 2.6. Reassign overlapped points and recover false noise.

False noise may exist in both outlier-clusters and pruned clusters, thus the corresponding clustered-index list and pruned-index list constructed in the Online phase are used for retrieving any such noise. A two-step recovery strategy is designed for iterating

the points contained in these lists. The first step works in the Offline phase, in which an attempt is made to merge each noise point into its nearest potential-cluster. Such merging is done based on distance: if the cluster radius remains below *epsilon* after adding the point to a potential cluster it is relabeled and added to this cluster, otherwise it is considered as noise.

The second step performs the same trail in the Post-processing phase after DBSCAN. This time, the remained noise points attempt to merge to their nearest offline-clusters formed by DBSCAN, by using the density threshold in DBSCAN. A successful merge would recover the corresponding false noise. The right panel in Figure 2.6 shows the first step as an example, since the second step follows the same process, only that potential-cluster is replaced by offline-cluster. Through the overlap reassignment and false noise recovery operations, the clusters after Post-processing will have correct and complete information of individual points.

2.3.3 Pruning with Real Time

The *tp* parameter in the original DenStream algorithm determines the time interval between pruning operations. A typical setting of this parameter is based on a "record count" threshold. This approach is suitable only under certain circumstances, for example when the records in a stream are received at regular time intervals. In practice, however, data streams may not always have a constant sampling rate, a situation that may lead to imbalanced pruning. An example of such a situation is shown in Figure 2.7, which depicts a data stream with a variable sampling rate that spans over several days. If pruning occurs every 10 points and the Online-Offline cycle is daily, then on the second day pruning will

not be invoked and all points will be used for clustering, ignoring the fact that the number of records during that day has declined. A more suitable approach for such a day would be to have fewer and smaller clusters since the amount of data records has decreased. In order to handle such situations GeoDenStream considers the time stamp of each record – and consequently the time difference between records – in order to invoke pruning using a time difference threshold. As a result, pruning is applied even when the number of records is decreased, causing less relevant records (i.e. temporally distant points) to be removed, and resulting in clustering results that better represent the actual data. In terms of implementation, the time interval for pruning in GeoDenStream can be constant or dynamic, set empirically or based on prior knowledge of the global information of the geographical data stream.

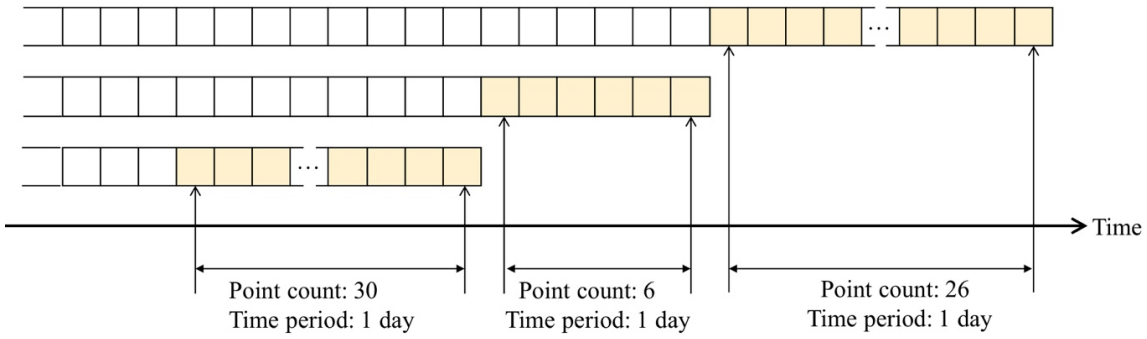

Figure 2.7. Time period and point count in pruning.

2.4 Implementation

The implementation of GeoDenStream was carried out based on the open source MOA project (Bifet et al., 2010), a widely used stream data mining framework (Kranen et al., 2012). Improvements in the basic four phases were achieved, and the Post-processing phase was established as an extensible Application Programming Interface (API). To link GeoDenStream with real geographical data stream, a prototype application supporting comma separated value (CSV) file is developed. There are four sets of parameters configurable in the proposed method:

(1) Stream configuration parameters, which include the geographical data stream in CSV format, the column index of timestamp, and the column indices of X and Y coordinates. Besides, column indices of an additional pair of X and Y coordinates can be incorporated as well. If this additional pair presents, the two pairs of coordinates are linked via behaviors with practical meaning, such as origin-destination of taxi trips, and source-sink of retweeting in Twitter.

(2) Time configuration parameters, which includes the starting time of the geographical data stream and the time interval information (i.e., type, value, and count) for Offline operation. The type of time interval can be year, month, week, day, hour, minute or second. The time interval value and count must be positive integers, and the count of interval determines the ending time for clustering.

(3) Clustering configuration parameters, including the ones in basic DenStream: *initial_points*, *min_points*, *epsilon*, *lambda*, *beta*, *mu*, and *offline*, and additional ones such as *pruning type* and *pruning value*. The *pruning type* parameter can be configured as Count

(pruning based on point sequence), Time (pruning based on a consistent time interval in seconds), and Dynamic (pruning based on evolving time intervals in seconds).

(4) Output configuration parameters, which include the directory for saving the clustering result files, the starting index of time interval for these output file names, an indicator of whether to export potential clusters, and another indicator of whether to apply GeoDenStream.

GeoDenStream is implemented in Java, and hosted at https://github.com/manqili/GeoDenStream. The code is compiled as a JAR package that is executable in Windows, Linux, and Mac OS platforms. An XML (extensible markup language) based document is designed for reusing this prototype. A sample configuration document is displayed in Figure 2.8.

```
<GeoDenStreamCase>
  <Stream>
    <FileName description="string: input csv file">c:/manqi/Boston/boston_export_forDEN.csv</FileName>
    <XColumnIndex description="integer: column index of X">6</XColumnIndex>
    <YColumnIndex description="integer: column index of Y">7</YColumnIndex>
    <TimeColumnIndex description="integer: column index of time">0</TimeColumnIndex>
    <ConnectedXColumnIndex description="integer: column index of the connected X">9</ConnectedXColumnIndex>
    <ConnectedYColumnIndex description="integer: column index of the connected Y">10</ConnectedYColumnIndex>
  </Stream>
  <Time>
    <StartTime description="string: Year-month-day hour:minute:second">2013-04-15 19:49:00</StartTime>
    <IntervalType description="string: Year|Month|Week|Day|Hour|Minute|Second">Minute</IntervalType>
    <IntervalValue description="integer: specifies time interval">10</IntervalValue>
    <IntervalCount description="integer: specifies time interval count">24*6=144</IntervalCount>
  </Time>
  <Cluster>
    <Initial_points description="integer: number of points used for initialization">100</Initial_points>
    <Min_Points description="integer: minimal number of points cluster contain">1</Min_Points>
    <Epsilon description="float: epsilon neighborhood">3.0</Epsilon>
    <Lambda description="float: lambda parameter for pruning">0.001</Lambda>
    <Beta description="float: beta parameter for pruning">0.2</Beta>
    <Mu description="float: mu parameter for pruning">1.0</Mu>
    <Offline description="float: epsilon*offline for offline DBSCAN">2.0</Offline>
    <PruningType description="string: Count|Time|Dynamic">Time</PruningType>
    <PruningValue description="Count: number; Time: number; Dynamic: file">900</PruningValue>
  </Cluster>
  <Output>
    <Directory description="string: output directory">c:/manqi/Manuscript/Boston/clusterdata/</Directory>
    <StartIntervalIndex description="integer: starting index of time interval">0</StartIntervalIndex>
    <OutputPotential description="integer: 0-No,1-Yes">0</OutputPotential>
    <ImproveProcessing description="integer: 0-No,1-Yes">1</ImproveProcessing>
  </Output>
</GeoDenStreamCase>
```

Figure 2.8. Configuration document of the prototype system.

2.5 Verification Using Synthetic Data

In order to verify that GeoDenStream produces results as expected, two types of synthetic datasets were created following the instructions in Hahsler et al. (2015), and implemented as inputs to GeoDenStream (datasets available at https://github.com/manqili/GeoDenStream). *DSD_BarsAndGaussians* and *DSD_Benchmark* were utilized to synthesize one static and one dynamic dataset,

respectively. *DSD_BarsAndGaussians* generated four clusters with varied densities that were composed of 5,500 points and 1% noise; two in uniformly distributed rectangular shapes and the other two were Gaissians clusters (Figure 2.9 (a)). *DSD_Benchmark* simulated two evolving clusters in a data stream with 5,000 points and 5% noise; one cluster moving from top left to bottom right and the other from bottom left to top right (Figure 2.10 (a)).

2.5.1 Visual Inspection

For both datasets, parameters in GeoDenStream were adjusted multiple times, in order to gain the best clustering results consistent with the reference. GeoDenStream for the static dataset was implemented with *epsilon* equal to 0.3 and *offline* equal to 3.0. Other parameters are specified in Figure 2.9. Figure 2.9 (b) demonstrates the distribution of the clusters identified by GeoDenStream, compared with the reference in Figure 2.9 (a). Visually we are suggested that the four clusters are successfully recognized, with well-defined boundaries and effectively removed noise, though some discrepancies still exist in the overlapped area of the two Gaissians clusters.

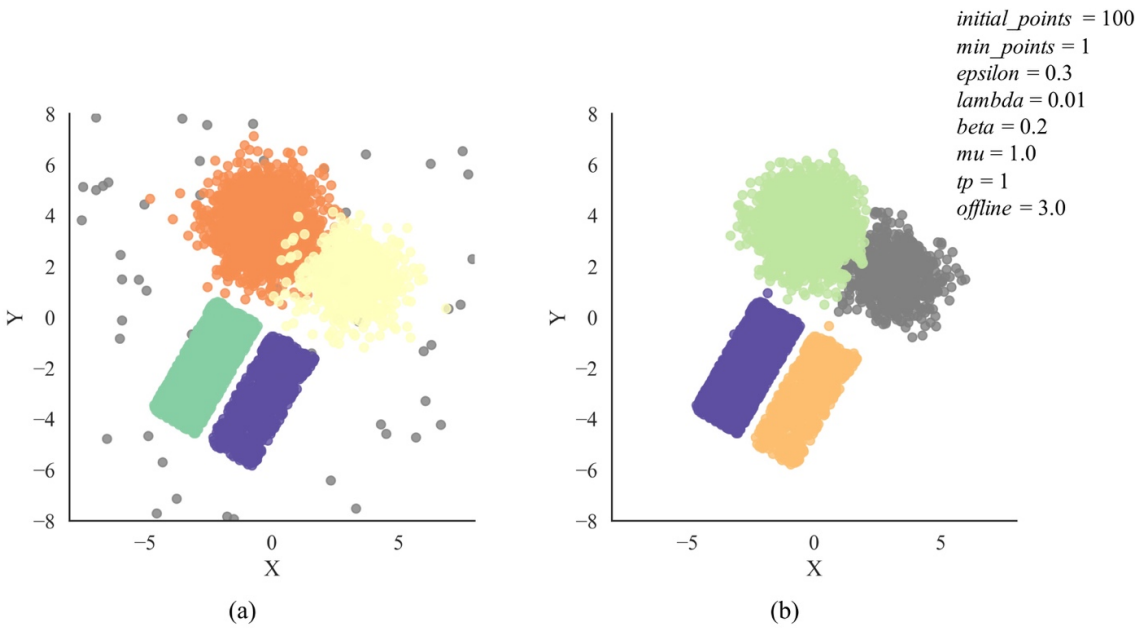

Figure 2.9. The static dataset (a) and clustering results using GeoDenStream (b) with parameters shown in the upper right corner.

Another combination of parameters was used for the evolving dataset and is listed in Figure 2.10. For this data stream, all points were divided into sequential subsets of 10 and imported to GeoDenStream step by step. Figure 2.10 (b) displays the cluster distributions determined by GeoDenStream, along with 'ground-truth' in Figure 2.10 (a), at selected time ticks. Except for the 6th time step, all clustering results showed high level of compliance with reference, and noise was cleaned effectively. Considering that the two clusters intersect at the 6th time step, it is almost impossible to separate them without manual interference. And hence evaluation of clustering results at this time step is skipped.

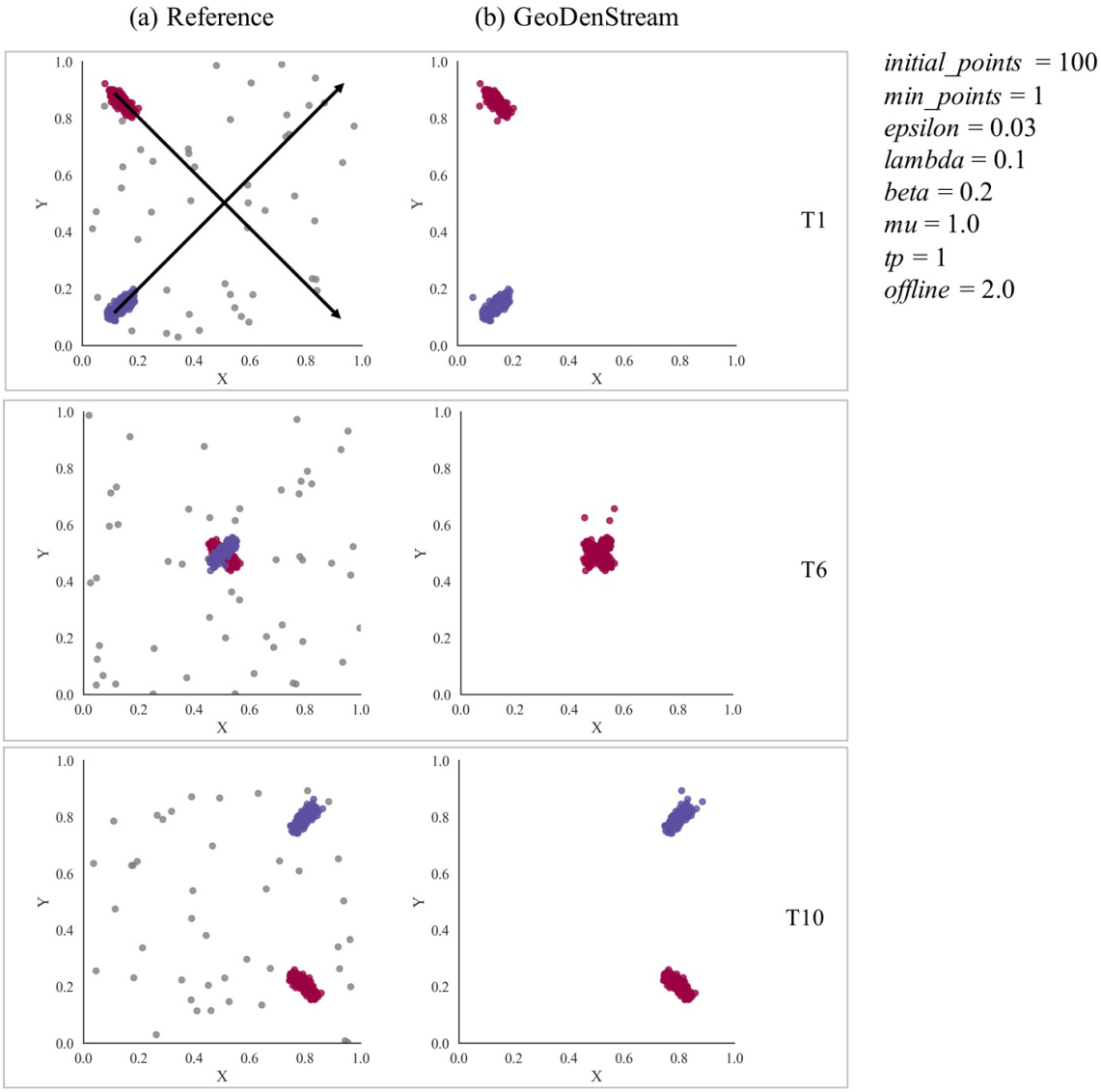

Figure 2.10. The evolving dataset (a) and clustering results using GeoDenStream (b) with parameters shown in the upper right corner.

2.5.2 Evaluation Metrics

Besides visual inspection, a series of metrics are provided to evaluate the performance of the proposed clustering method mathematically. These evaluation metrics include: adjusted Rand index and Silhouette Coefficient for overall assessment; and precision, recall, and F1 score for evaluating single clusters. Overall assessment tests if the clustering method defines separations similar to the reference, and if members belong to the same cluster are more similar than those in different clusters. The adjusted Rand index is suitable for the first purpose since it is a popular similarity measure that compares two sets of assignments (Vinh et al., 2010). Silhouette Coefficient fits for the second purpose, by measuring the mean distance between a point and all other points in the same class, as well as points in the next nearest cluster. A higher Silhouette Coefficient score indicates better defined clusters (Rousseeuw, 1987). Precision looks for how many selected points are relevant, recall checks how many relevant points are selected, and F1 score is the harmonic mean of them (Powers, 2011).

There were four clusters identified from the static data by GeoDenStream, same as that in the reference records. Evaluation measures of this dataset are summarized in Table 2.3. The high values of adjusted Rand index imply favorable clustering results comparing to the 'ground truth,' and the comparability of Silhouette Coefficients of GeoDenStream clustering results and of reference shows the agreement on cluster structures of the two assignments.

Precision, recall, and F1 score were mainly used for assessing the method's ability to handle noise. The relatively high recall and low precision values of cluster 0 suggest that

most of the real noise was detected by the method, however a higher percentage of non-noise points were also included. Referring to Figure 2.9 (a), I observed that due to the random locations these noise points are placed, some of them are close enough to the clusters to be exempt from noise. Also, the overlapped area of the two Guassian clusters indicates the difficulties of setting distance-based parameters, and further results in problems in noise judgement. Thus, it is suggested that the performance of noise detection does not solely depend on the GeoDenStream method.

Table 2.3. Evaluation measures of GeoDenStream using the static dataset; cluster ID = 0 means noise.

Adjusted Rand Index					0.960					
Silhouette Coefficient (reference)					0.831 (0.819)					
	Predicted									
Reference	Cluster ID	0	1	2	3	4	Total	Precision	Recall	F1 score
	0	**42**	8	2	3	3	58	0.247	0.724	0.368
	1	30	**1954**	17	1	0	2002	0.985	0.976	0.981
	2	45	21	**586**	0	1	653	0.969	0.897	0.932
	3	0	0	0	**2099**	0	2099	0.998	1	0.999
	4	53	0	0	0	**635**	688	0.994	0.923	0.957
	Total	170	1983	605	2103	639	5500	0.981	0.967	0.972

Evaluation measures of the evolving data were calculated at each time tick (Table 2.4). As stated in section 2.5.1, the 6th time step was skipped for analysis. For all the other

clustering results, measures for overall assessment illustrate the strong agreement between GeoDenStream clustering results and the reference. F1 score for the noise class proves the superior capability of handling noise in this dataset.

Table 2.4. Evaluation measures of GeoDenStream using the evolving dataset.

Time	Adjusted Rand Index	Silhouette Coefficient (reference)	F1-score of noise
1	0.987	0.839 (0.826)	0.968
2	0.987	0.819 (0.806)	0.966
3	0.996	0.814 (0.810)	0.987
4	0.986	0.762 (0.753)	0.965
5	0.976	0.600 (0.595)	0.981
6	0.251	0.785 (0.034)	0.96
7	0.991	0.638 (0.638)	0.979
8	0.974	0.758 (0.741)	0.935
9	0.978	0.818 (0.803)	0.935
10	0.986	0.831 (0.819)	0.965

2.6 Case Studies

Two case studies were conducted in order to examine the utility of GeoDenStream for clustering spatiotemporal data from social media streams. In both cases studies Twitter, a popular social media platform, served as the data source. In particular, two Twitter data streams were collected: a first set including 649,663 tweets about the Boston Marathon bombing in 2013; and a second set including 984,967 tweets was about the Zika virus epidemics in 2015. Both datasets were collected using a worldwide keyword-based search using Twitter's streaming API. Table 2.5 provides a summary of these two datasets. It should be noted that in both case studies both precisely (GPS based) and imprecisely

(toponym based) geolocated tweets were used. This is particularly important in the context of this work as imprecisely geolocated tweets often result in multiple tweets with identical spatial coordinates, thus allowing us to evaluate GeoDenStream's ability to handle overlapping points. In addition, it is worth noting that the two datasets span over different temporal intervals, thus enabling us to demonstrate how the algorithm can be applied at different time granularities (i.e. hours versus days).

Table 2.5. A summary of the used Twitter datasets.

	Boston Bombing	Zika
Spatial Extent	Worldwide	
Geo-reference information	X, Y coordinates of tweets and their retweets	
Geo-referenced tweet count	649,663	984,967
First Timestamp (UTC)	2013-04-15 19:49:06	2015-12-12 00:00:00
Last Timestamp (UTC)	2013-04-16 19:49:06	2016-03-05 00:00:00
Duration of Time	24 hours	84 days

2.6.1 The Analysis Process

The analysis of the two datasets was performed in two steps, namely GeoDenStream clustering (step 1) and spatiotemporal analysis (step 2), as shown in Figure 2.11. In the clustering step each dataset was first analyzed to detect whether each record (tweet) is an original message or a retweet. When a retweet was detected the tweet ID of the original message (if exists in the dataset) was also stored in order to enable analyzing a corresponding retweet network in step 2. While the objective of step 1 is to demonstrate

the capability of GeoDenStream to generate spatiotemporal clusters, the objective of step 2 is to demonstrate the utility of the information that is captured during the clustering for subsequent analysis based on the derived clusters.

In order to perform the clustering in step 1 each data set was used to simulate a real-world data stream. Specifically, each dataset was ordered temporally using the timestamp of each record, and then each ordered dataset was progressively fed into the GeoDenStream clustering algorithm. During the clustering process information about the record-wise composition of each cluster was captured. This information is used later in step 2 in order to explore the retweet network structure both within and between spatiotemporal clusters.

Analysis in step 2 consisted of two types of analyses: an examination of the properties of the retweet network within a cluster at each analysis time step, and an examination of the properties of the retweet network between clusters at each analysis time step. Within each cluster several commonly used network- and node-level properties were calculated: the number of nodes and edges, the overall network density, the size of the largest network component as a percentage of the whole network, and the average closeness and eigenvector centralities. In addition, a worldwide retweet network between different clusters was constructed and visualized on the map at each time step. All analyses were performed using Java(TM) version 1.8.0_102 on system with 4GB and an Intel(R) Core(TM) i5-5250U CPU, running Windows 10.

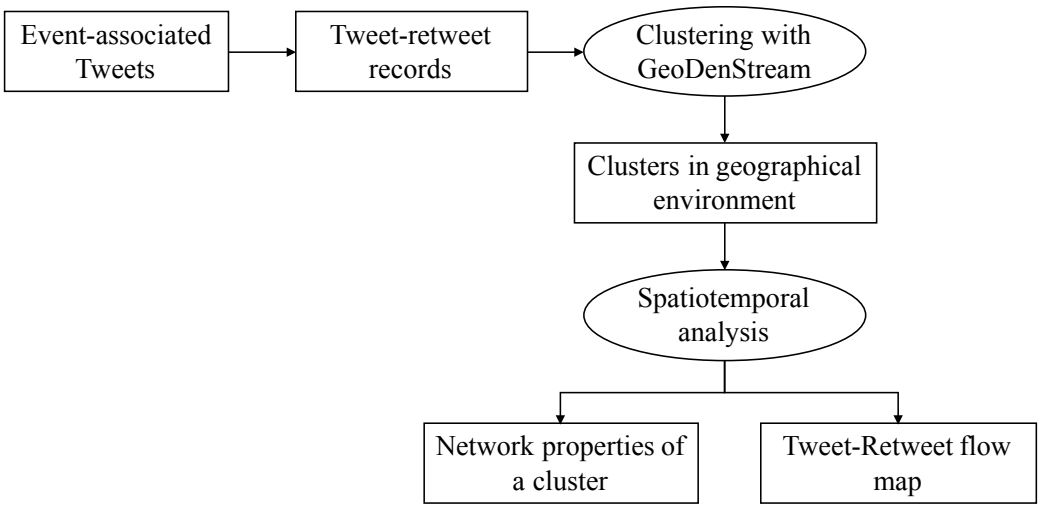

Figure 2.11. Spatiotemporal analysis based on GenDenStream results.

2.6.2 Results

2.6.2.1 Clustering with Overlapping Points

The derivation of the spatiotemporal clusters was carried out using GeoDenStream in each case study. In the 2013 Boston Bombing case study, the Offline and Post-processing phases were conducted hourly, resulting in 24 sets of clustering results. In the 2015 Zika case study the Offline and Post-processing phase were conducted daily, resulted in 84 sets of clustering results. The resulting cluster and point counts for each time interval in the Boston Bombing and the Zika case studies are presented in Figure 2.12 (a) and (b) respectively. Figure 2.13 shows examples of the clustering results from GeoDenStream using the time interval in each case study that corresponds to the highest number of clusters (11^{th} hour for the Boston Bombing dataset and 53^{rd} day for the Zika dataset).

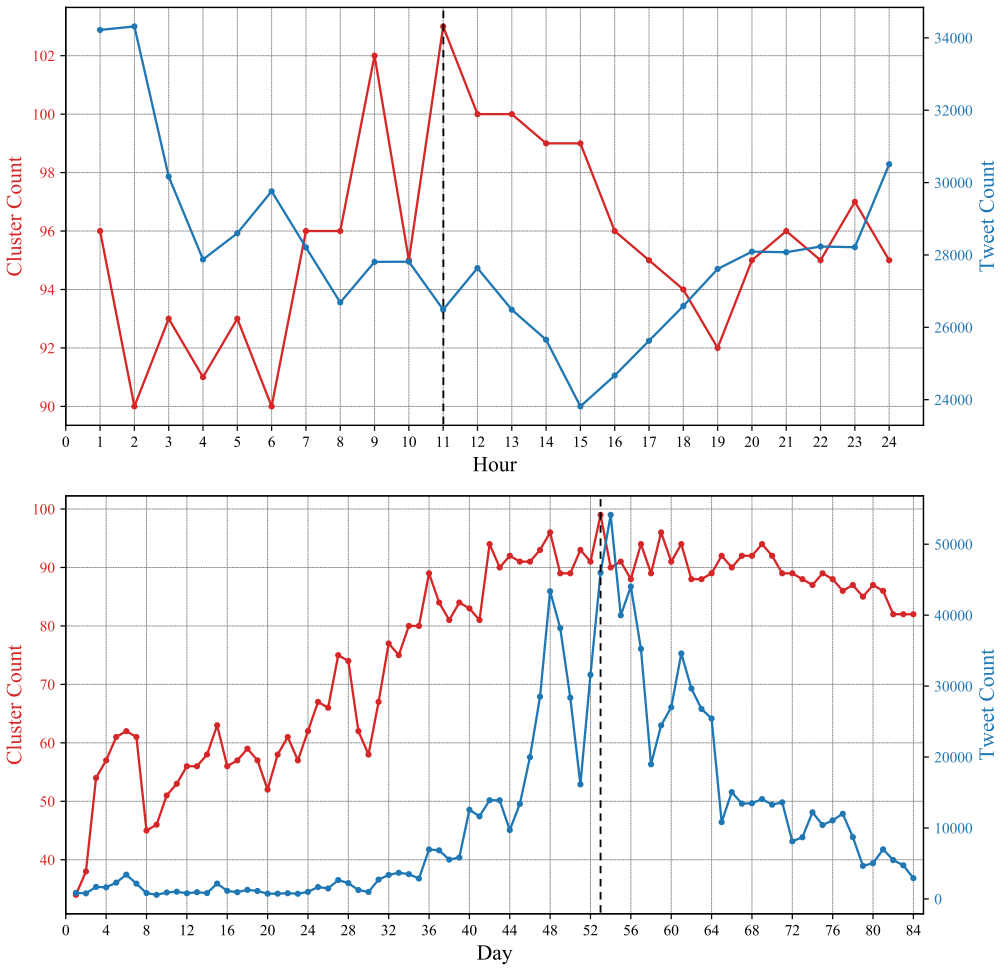

Figure 2.12. Cluster and point count of (a) Boston Bombing, and (b) Zika.

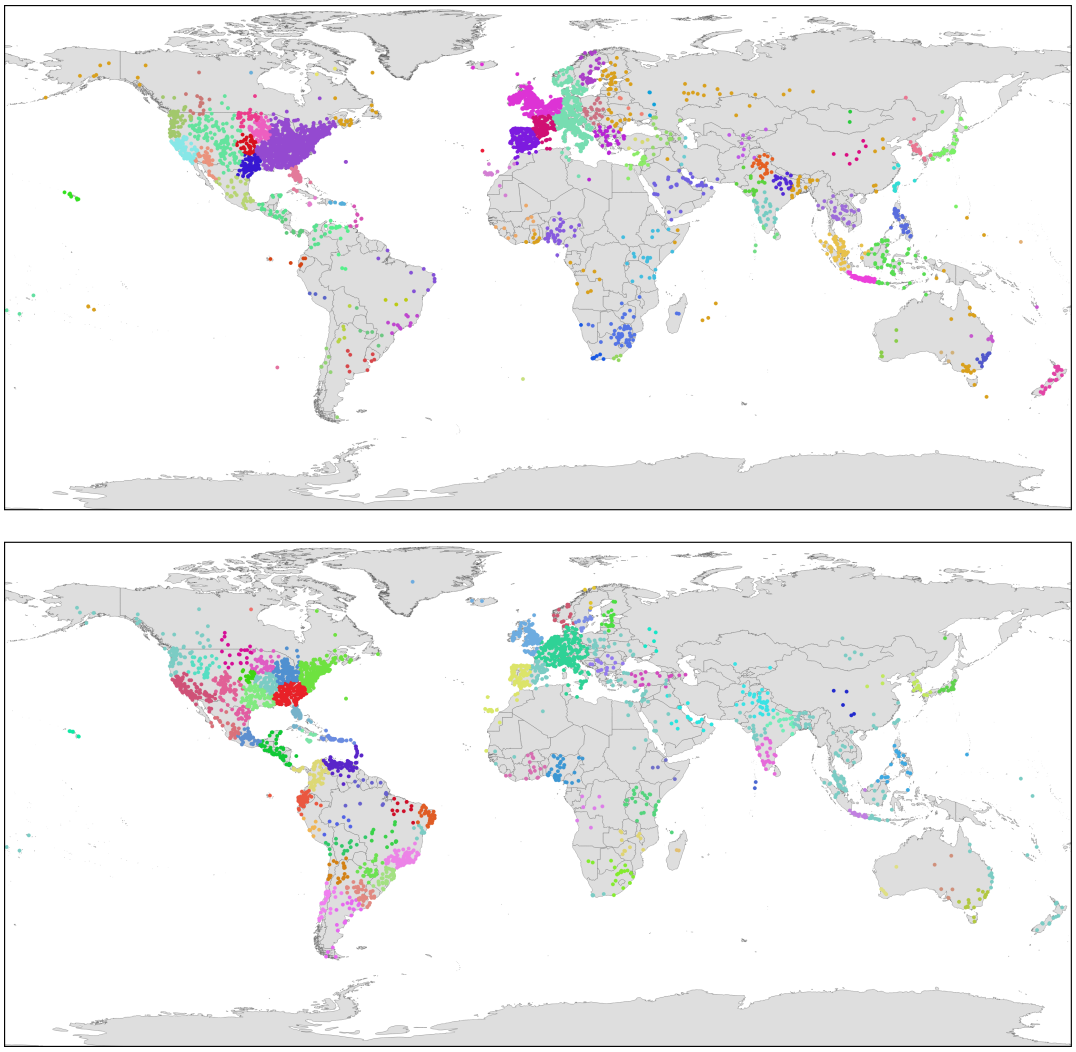

Figure 2.13. Sample results of the GeoDenStream clustering results for the time interval with the highest number of clusters: (a) Boston Bombing in the 11th hour and (b) Zika on the 53rd day; each color represents a cluster.

In order to further examine these results with respect to GeoDenStream's ability to handle instances of overlapping points and false noise, the clustering results of such points were examined by applying both the original DenStream algorithm (Cao et al., 2006) and

the proposed GeoDenStream algorithm. An example of the clustering results obtained from both algorithms in the Zika case study is shown in Figure 2.14. The red arrow in Figure 2.14 (a) points at two overlapping points, which were assigned to two different clusters (in this case clusters 17 and 31) by DenStream, disregarding their location overlap. In GeoDenStream, however, these two points were reassigned to the same cluster after the Offline phase (as described in Section 2.3.2).

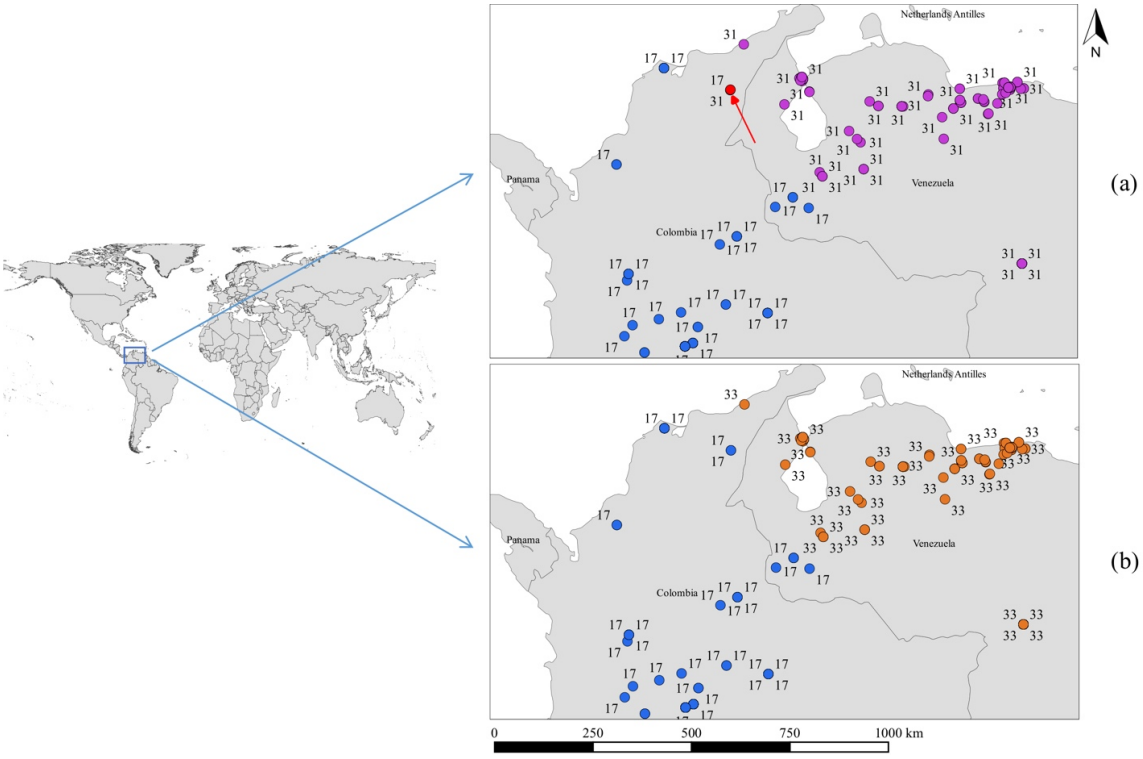

Figure 2.14. Point overlap issues with basic DenStream (a) and GeoDenStream (b) using the 3rd day of Zika dataset. Numbers next to points indicate cluster numbers, the red arrow in (a) indicates the location of overlapping points in the data.

An example of GeoDenStream's ability to handle false noise in the Offline and Post-processing phases compared to the classic DenStream algorithm in the Zika dataset is shown in Figure 2.15. In particular, Figure 2.15 (a) shows the results obtained from DenStream. The red arrows in this figure point at three noise points, among which the one in Norway is a true noise point and the other two are false noise points. While in DenStream these three points are all regarded as noise, in the GeoDenStream algorithm the false noise points are merged into their respective nearest cluster, through additional two steps in the Offline and Post-processing phases, as shown in Figure 2.15 (b).

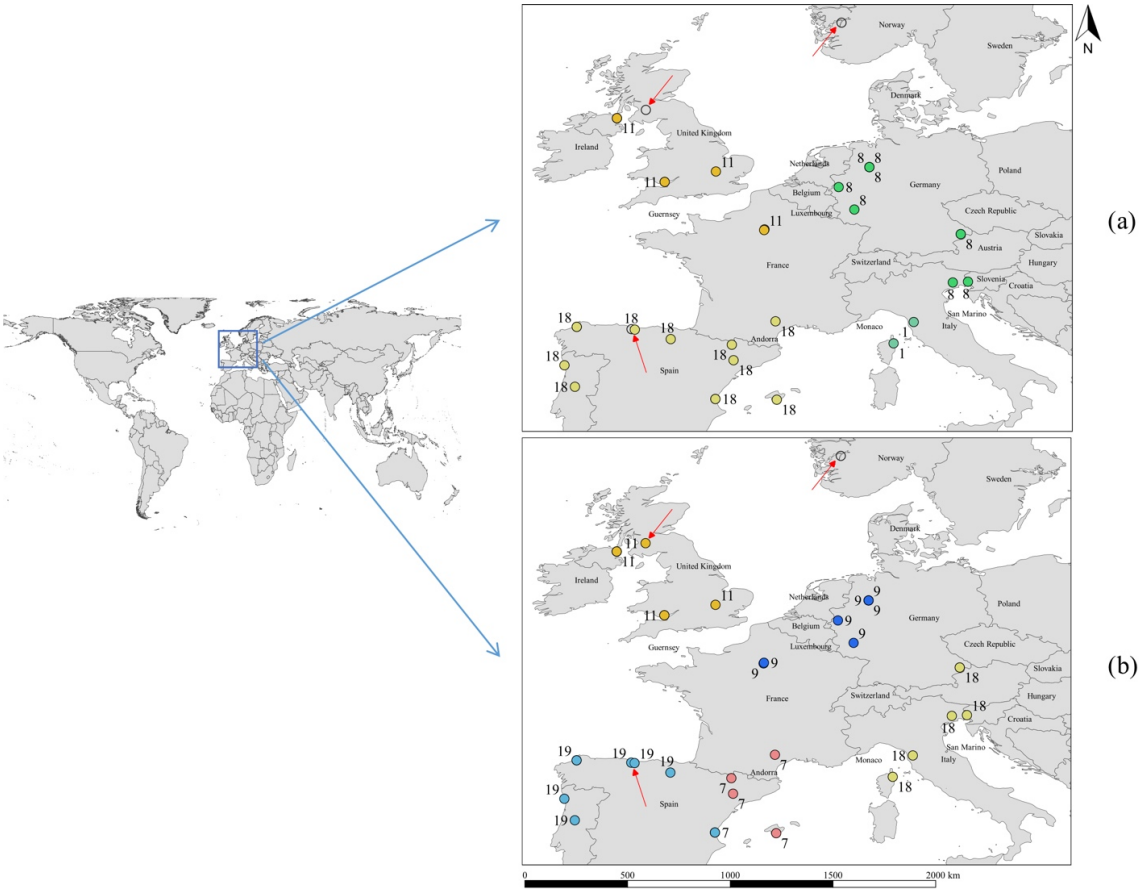

Figure 2.15. False noise issues with basic DenStream (a) and GeoDenStream (b) using the 1st day of Zika dataset.

2.6.2.2 Memory Usage in GeoDenStream

In order to evaluate the benefit of indexing (Section 2.3.1), in the GeoDenStream implementation memory usage was monitored during the clustering of the two case study datasets. In particular, two processes were run: a clustering process without the proposed indexing, and another with it. Figure 2.16 depicts the results of these two processes for both the Boston Bombing and the Zika virus datasets. As this figure shows, in the Boston

Bombing dataset the amount of memory used with indexing was always smaller than the case without indexing in every time interval. On average, hourly memory usage without indexing was 272.81 MB while with indexing it was about 158.44 MB, a reduction of 41.9% of the hourly memory usage. Similar results were obtained in the Zika dataset: while the average daily memory usage without indexing was 134.5MB, it reduced to about 68 MB with indexing, a reduction of 49.4% of the average daily memory usage. These reductions show GeoDenStream's ability to substantially lessen its memory usage compared to DenStream.

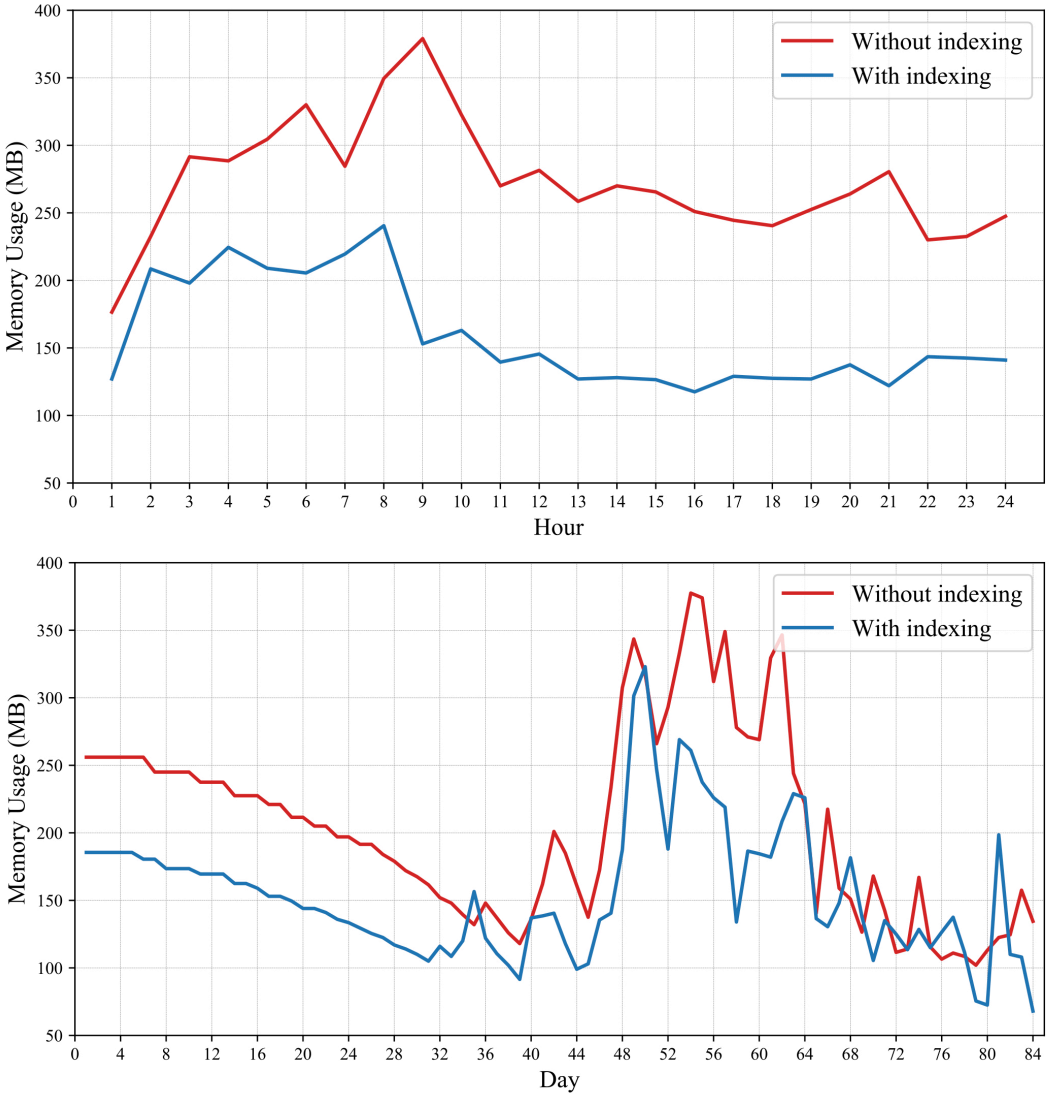

Figure 2.16. Memory usage with and without indexing stream points using (a) Boston Bombing dataset and (b) Zika dataset.

2.6.2.3 Network Analysis within a Cluster

In addition to the clustering results from GeoDenStream, it may be of interest in some application domains to analyze the relationships between clusters. For example, when analyzing geotagged social media streams, it is often useful to consider both where clusters

of users are located as well as how users communicate both within and across clusters. Such geosocial analysis is important as it can provide important insights into the exchange and flow of information over space and time, which may ultimately affect human behavior.

In order to demonstrate how such analysis can be applied using GeoDenStream for the analysis of the two case study datasets, a representative cluster was selected in each case study, and an analysis of the retweet network within each cluster was carried out. Specifically, in each case study the cluster with the highest volume of accumulated retweets was selected, resulting in one cluster located in Boston and its surrounding areas (for the Boston bombing case) and one cluster covering Venezuela and its neighboring countries (for the Zika virus case). Then, for each cluster, a set of network-level (node and edge counts, density, and the proportion of the largest component of the network) and node-level (closeness and eigenvector centralities) measures were calculated over the analysis period of each case study. These measures were selected as they are often used to analyze and characterize social networks (Boccaletti et al., 2014).

An overview of the results of this analysis is shown in Figure 2.17. The line plots in the upper half of this figure show the network-level measures. The boxplots in the lower half of Figure 2.17 show the distribution of the node-level centrality measures at each time step, and how this data distribution varies over time. As can be seen from this figure, GeoDenStream's ability to consistently maintain information about clusters throughout the clustering process enables the network analyses to successfully capture the temporal dynamics of the network over time.

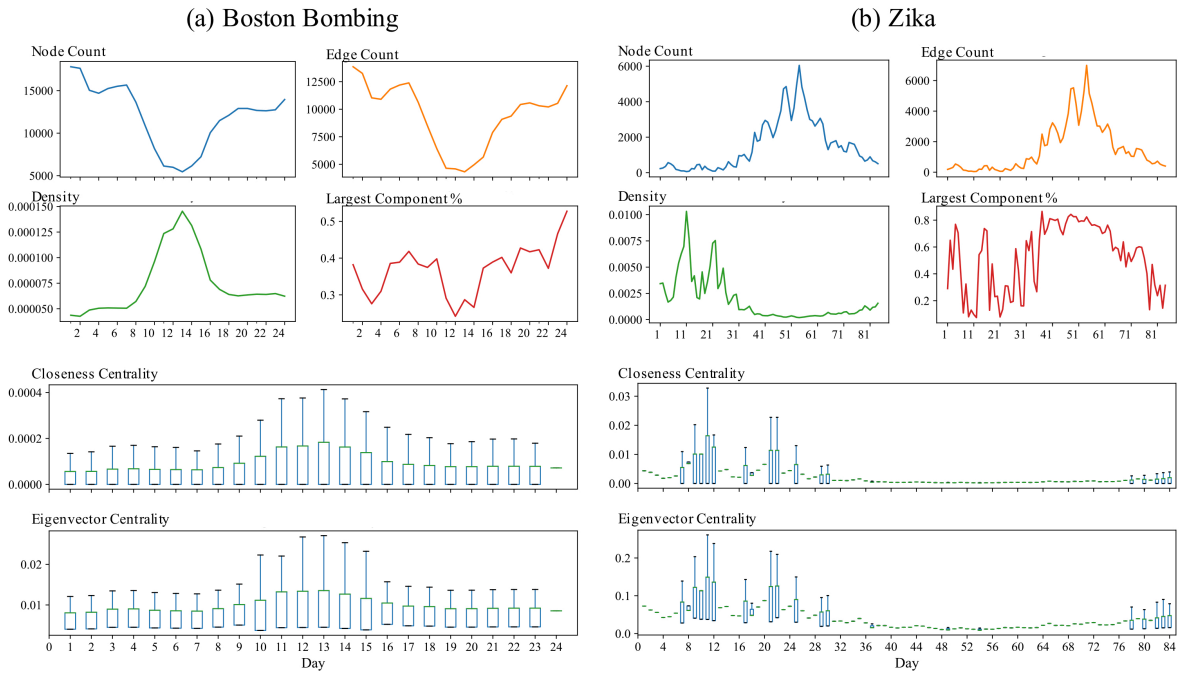

Figure 2.17. Network Properties of representative clusters of (a) Boston Bombing and (b) Zika.

2.6.2.4 Network Analysis between Clusters

Following the network analysis within a cluster, analysis of the network components between clusters was carried out. Here, information flow between users in different clusters was analyzed and visualized on a world map to show the overall global information flow. Figure 2.18 shows an example of the results of such analysis, in which information flows between clusters are shown as arcs in the counter-clockwise direction from the users authoring tweets to users retweeting them, and within-cluster flows are represented by self arcs. The color of the arcs defined in the color bar indicates the communication flow frequency, and cluster IDs are labeled as numbers on the map. For visual clarity, Figure

2.18 shows only the top 30 arcs from the 75th percentile of all arcs in the network in terms of their retweeting (or information flow) volume.

The utility of these flow maps is in providing insights about how information propagates globally. Specifically, they reveal the locations of clusters that serve as information sources or sinks. For example, this analysis reveals the frequent information exchange between the United States and the West Europe in the case of the Boston Bombing (Figures 2.18 (a) and (b)), and shows Venezuela and Brazil as major information source to the United States and West Europe in the case of the Zika virus (Figures 2.18 (c) and (d)). Such analysis can also assist in identifying key changes in the flow of information between regions. For instance, in the case of the Zika virus Venezuela and Brazil acted as the major information source to other clusters on the first day but later acted as self sources and sinks on day 42 (Figure 2.18 (c) and (d)).

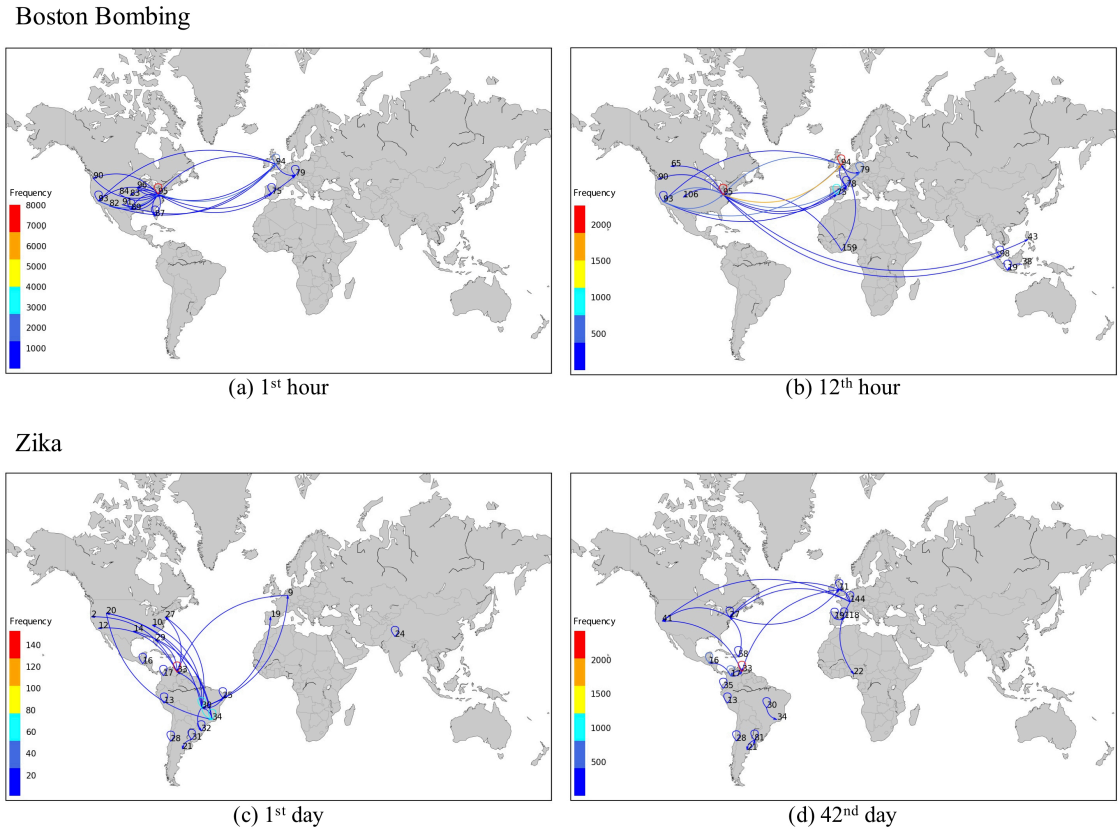

Figure 2.18. Retweet flow maps in the Boston Bombing case study (top row) and the Zika virus case study (bottom row); cluster IDs are labeled as numbers, flow frequency is indicated by the color bar, and flow direction is represented by the counter-clockwise arcs.

2.7 Discussion

GeoDenStream is a novel tool for clustering spatiotemporal data streams. Building on DenStream, this tool is particularly suitable for analyzing entity-based geographical data streams such as social media data due to three unique characteristics: its ability to track and

maintain information about the identity and composition of clusters over time and space, its ability to handle spatially overlapping data points, and its improved ability to handle noise. These capabilities are achieved primarily through the integration of an indexing scheme into the clustering process, which also substantially reduces the memory requirement of the clustering process. In order to demonstrate and evaluate the utility of this tool both synthetic and real-world stream data was used. Given its performance and characteristics, I envision that it could be broadly used to identify, track, and infer entity-driven activity from various data streams, such as geotagged social media, location-aware mobile devices, urban monitoring networks, and vehicle tracking.

Currently, there exist several well-established open source software libraries that offer an implementation of DenStream: the MOA Java package (Bifet et al., 2010), the streamMOA R package (Hahsler et al., 2015), and the OutlierDenStream Python package (2018). Their basic characteristics are provided in Table 2.6. MOA package supports Initializing, Online, Pruning, and Offline phases. Because information of individual points in micro-clusters or points pruned is not recorded, it's impossible to identify noise points or to access points contained in a cluster with MOA. streamMOA is an R wrapper of MOA package. The process of conducting Initializing, Online, and Pruning phases are consistent with that in MOA package, with one difference that the Offline phase is based on Reachability or Hierarchical clustering method that are defined in the stream package in R (Hahsler et al., 2017). In the Offline phase, all points including noise are assigned with a cluster ID, which brings about inaccuracy of the clustering results. In OutlierDenStream package, the process of Initializing, Online, and Pruning phases is similar to that in MOA

package. An advance of OutlierDenStream is that real time can be used as timestamps. However, it does not have Offline phase, and its clustering results are only presented as potential-clusters and outlier-clusters without aggregation.

Table 2.6. Popular implementations of DenStream clustering method.

Package	Language	Pros	Cons
MOA	Java	Support Initializing, Online, Pruning, and Offline phases.	Does not provide access to relations between points and clusters.
streamMOA	R	Support Initializing, Online, Pruning, and Offline phases.	Relations between points and clusters are built without noise removal.
OutlierDenStream	Python	Support Initializing, Online, and Pruning phases; real time can be used for pruning.	Does not support Offline phase, so there is no access to relations between points and clusters.

Table 2.7 summarizes how GeoDenStream compares to these packages in terms of cluster processing, the point pruning strategy employed, memory usage, and the handling of overlapping points and noise. As can be seen from this comparison, GeoDenStream offers some unique capabilities that are not available in these commonly used libraries. Moreover, GeoDenStream enables obtaining and tracking the mapping between data records and clusters throughout the clustering process.

Table 2.7. Comparison of different DenStream clustering methods.

Functionality		MOA	streamMOA	OutlierDenStream	GeoDenStream
Phase	Initializing	√	√	√	√
	Online	√	√	√	√
	Pruning	√	√	√	√
	Offline	√	√		√
	Post-processing				√
Pruning strategy	Point count	√	√	√	√
	Constant time			√	√
	Dynamic time				√
Memory usage reduction		√		√	√
Data overlap handling					√
Noise removal		√		√	√
False noise recovery					√
Data records & clusters mapping					√

Following the work presented in this chapter, there are several possible avenues for further expanding and improving GeoDenStream. In particular, I envision that future work will focus on three primary areas: improved indexing, context-aware clustering, and parallel acceleration. In terms of indexing, improvements to the indexing scheme that is used for exploring neighborhood relationships (e.g., CPM (Mouratidis et al., 2005)) could be integrated into GeoDenStream to support more efficient nearest-neighbor searches for new stream data points. In terms of context-aware clustering, improvements to GeoDenStream's processing phases could be made in order to introduce non-geographic

attributes (e.g., social media messaging topic similarity or social affinity) that can be particularly important to the formation of meaningful clusters. As for parallel acceleration, the computational efficiency of spatial clustering can be enhanced by parallel CPU and GPU computing (Chen et al., 2018b; Skála and Kolingerová, 2011), which is extraordinarily meaningful when the data size grows large, since millions or even billions of data points are common in the context of Big Data.

3 SPATIOTEMPORAL ANALYSIS OF INFORMATION DIFFUSION IN EVENT DISCUSSION OVER TWITTER

Abstract

Understanding the dynamics of information diffusion in social networks contributes to a wide range of social studies. Among all kinds of social networks, Twitter is of my particular interest due to its richness, availability, and popularity. This research aims at exploring the spatiotemporal patterns of information diffusion in discussions about real-world events over Twitter. It applies information source-to-sink flow analysis based on the multi-temporal similarity measure, and contributes to our understanding of how a cyber population reacts to public health-related issues via case studies on epidemics. Results suggest that (1) geographic and temporal scales are worth exploring, due to their influence on the clustering results and on the information diffusion patterns discovered; (2) the information flow analysis along with similarity measures were able to capture the information diffusion patterns, regarding direction, volume, locations of its source and sink, as well as its temporal evolvement; and (3) several influential factors on information diffusion were uncovered, including participants' distribution and activeness, geographical scale, geographical location, geography-driven homophily, and time of events progress.

3.1 Introduction

People are acting as sensors as they participate in the cyber activities and interact with each other as well as the cyber environment. The emergence of various social media services and platforms, such as Twitter, Facebook, and Flickr, has fostered increasing contributions from the individuals to the generation and dissemination of public information. Though having a large volume, high volume, variety, velocity, and veracity, the contributions from these human sensors in social media are considered valuable yet limited to individual perception (Croitoru et al., 2014).

Twitter, a popular social media platform that claimed 321 million monthly users in February 2019 (Shaban, 2019), has fostered increasing participation in short message sharing. Thus, one of the highlighted usage of Twitter is information propagation, for which retweeting is the key mechanism (Suh et al., 2010). Retweeting is the behavior of sharing a tweet written by another user with one's own followers. A retweet can be generated in one of two ways. First, one can retweet with one-click on the 'Retweet' icon. Second, one can also add comments before retweeting, making it a "Quote Tweet." Retweeting has been extensively studied. For instance, Boyd et al. (2010) investigated retweet activity about what, why, and how people retweet. Zarrella (2009) proposed that the characteristics between retweets and normal tweets vary fundamentally. Retweetabilility has been a focal topic since it is closely related to the efficiency of information sharing (Lotan et al., 2011), users' popularity and influence (Bakshy et al., 2011; Cha et al., 2010), community identification (Croitoru et al., 2015), and event detection (Atefeh and Khreich, 2015). Targeting on a certain event, the retweeting activities

among Twitter users characterize the diffusion of relevant information, which helps to understand the dynamics of this event from the lens of the general public.

Nowadays, there is a growing awareness of the value of geo-referenced contents in Twitter, which has inspired the emergence of research stressing the geographic environment, since a variety of Twitter activities are deeply embedded in it (Ferrara et al., 2013; Java et al., 2007; Kulshrestha et al., 2012; Pruthi et al., 2015). Previous studies about Twitter with explicit geo-references span a wide range of research interests, including the geographic distribution in the Twittersphere (Java et al., 2007), the substantial impact of geography on user connections (Kulshrestha et al. 2012), relation between trending topics and geography (Ferrara et al., 2013), spatiotemporal extent of local and global events (Pruthi et al., 2015), and geographic distribution of sentiment (Agarwal et al., 2018).

Though geographic influence on Twitter activities has been well-investigated, explorations of geographical influence on information diffusion over Twitter is still in demand. Previous studies on information diffusion considering geography mainly focused on the changes of reached location and coverage area (Kwon et al., 2015; Pruthi et al., 2015; Puri et al., 2018). However, interactions within the participants at different locations at a specific time is unknown. I identify this issue as information flow, linking the source (i.e., from where information is generated) and sink (i.e., to where information is spread) of a diffusion process. In limited number of studies involving information flow in Twitter, it was usually utilized for visualization (Croitoru et al., 2015; Lotan, 2011; Mishori et al., 2014); while mining meaningful patterns behind it is overlooked, especially in the geographical environment.

To fill the gap of characterizing information diffusion over Twitter in the blended cyber and geographic space, this chapter targets the source-to-sink information flow trends, as well as the information source and sink formed by the flow. In this way, interpretation of the direction, volume, spatial distribution, and temporal evolvement of information diffusion can be generalized. This work contributes to depicting the internal mechanism of information diffusion, surfacing the essence veiled in the mass data, and facilitating decisions and further applications in an accurate, responsive, and flexible manner. The remainder of this chapter is organized as follows: Section 3.2 describes two case studies used in this chapter; Section 3.3 presents the designed methods; upon the results presented in Section 3.4, discussions are performed in Section 3.5; Section 3.6 concludes this part of work.

3.2 Case Studies

In this study, the targeted information is substantiated as discussions associated with real-world events in Twitter, and is propagated by the retweeting behavior. Two public health emergencies of international concern are chosen as case studies: one is about Zika virus disease, and the other concerns Ebola virus disease. The most recent outbreak of Zika was first reported in 2015, starting from Brazil, and then spread to other South American areas (Kindhauser et al., 2016). Since October 2015, a growing number of countries experienced Zika virus outbreaks, roughly covering the whole South America and lower regions of North America (WHO, 2016). As for Ebola, the West African Ebola virus epidemic during 2013-2016 was the most widespread and complex outbreak since its discovery in 1976. It was first officially declared an outbreak in March 2014, peaked in October, and then started

to decline gradually in response to substantial international assistance. It has extended from countries in West Africa to Italy, Mali, Nigeria, Senegal, Spain, the United Kingdom, and the United States (WHO, 2018). I chose to study these two epidemics because of the worldwide concern they have raised when prevalent, as well as the continued attention from the public regarding outbreak responses such as understanding the role of social media in an outbreak of disease (Jacobsen et al., 2016; Stefanidis et al., 2017) and reducing the spread and severity of future diseases (Hoffman and Silverberg, 2018).

Twitter data capturing the participation of the global community in these events are harvested by the GeoSocial Gauge system (Croitoru et al., 2013). This system fetches the event-related tweets through the Twitter application program interface (API) using a worldwide keyword-based search. Keywords used for collecting the two datasets are "zika, chikungunya, zikv" and "ebola", respectively. Meanwhile, the associated metadata are also captured. In the metadata, the combination of timestamp, location, author, and content corresponds to the key elements in the information diffusion process: when, where, who, and what. A document-oriented NoSQL database, MongoDB is used for the Twitter data storage and management.

A total of about 6.2 million and 52 million tweets covering the whole world in a duration of 84 and 120 days were collected for Zika and Ebola events, respectively. Table 3.1 summarizes the data basics of the two datasets. In the Zika dataset, 3,220,485 (51.53%) of the tweets contain location information. Within these georeferenced tweets, 1,350, 281 (21.61%) are retweeted from other users. And in these georeferenced retweets, 984,967 (15.76%) of them provide traceable location information of their sources in the whole

dataset. The subsequent analyses are performed upon this set of data. Same subset was acquired for the Ebola case, which consists of 5,150,085 georeferenced retweets with georeferenced source, 9.85% of the whole Ebola dataset. All subsequent analyses are based on a daily interval.

Coordinates of the georeferenced tweets are from three types of sources, listed from the most to least reliable: coordinates provided by the user (e.g., GPS, cell tower triangulation, etc.), coordinates inferred by Twitter from IP address, and coordinates from the location toponym in a user's profile. Ideally, the first type would be the most desirable; however by querying the database storing Zika and Ebola datasets, I found that among all georeferenced records from any type of sources (i.e., 984,967 and 5,150,085 tweets in Zika and Ebola datasets, respectively), there are only 12,115 (1.23%) and 167,893 (3.26%) containing the exact coordinate information. Considering this practical situation, coordinates from Twitter (20,783 (2.11%) and 33,990 (0.66%)) and from location toponym (952,069 (96.66%) and 4,948,202 (96.08%)) are also considered as valid coordinate information in these two case studies.

Table 3.1. Summary of data basics; subsequent analysis is based on the underlined datasets.

	Zika		Ebola	
Whole	6,249,626	100%	52,298,510	100%
Georeferenced tweets	3,220,485	51.53%	19,182,533	36.68%
Georeferenced retweets	1,350,281	21.61%	9,409,318	17.99%
Georeferenced retweets with georeferenced source	984,967	15.76%	5,150,085	9.85%
Coordinates from GPS	12,115	0.2%	167,893	0.3%
Spatial extent	Worldwide			
Temporal resolution	Daily			
Starting date	2015-12-12		2014-08-21	
Ending date	2016-03-05		2014-12-18	
Time duration	84 days		120 days	

3.3 Methods

This study consists of three main parts: stream clustering, information flow construction, and spatiotemporal analysis. An overall workflow is summarized in Figure 3.1. For a dataset, first the raw Twitter records are harvested and organized in a MongoDB database. Then applying a series of filtering criteria and preprocessing steps, I obtain a set of useful records—the underlined subset of retweets in Table 3.1. Tracing each retweet's source among the georeferenced tweets, I obtain its location information and assign it to the corresponding retweet. Hence in this subset, each retweet contains two pairs of coordinates, one pair of itself and the other of its source tweet.

For analyzing data with large volume and high velocity, aggregating individuals into meaningful groups is essential due to practicability. Therefore, on the subset, retweets

are clustered by a stream clustering method using their location information on daily basis, and then the cluster membership of their source tweets is also identified using their location information. Thus, the resulting clusters contain data points representing both retweets and source tweets, and thus have spatiotemporal features induced from them. For example the location of a spatiotemporal cluster is defined as the center of the smallest convex shape enclosing all its containing data points. At the cluster level, I aggregate the tweet-to-retweet occurrences, which construct information flow among the clusters.

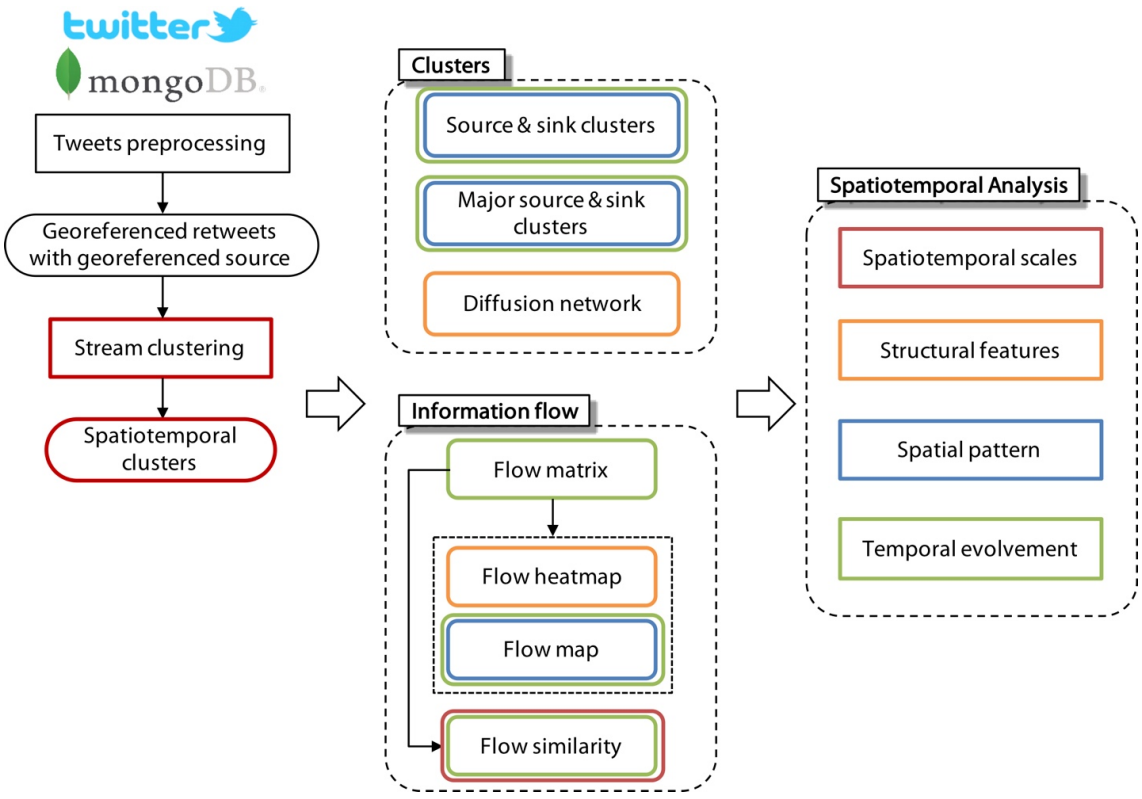

Figure 3.1. Workflow of spatiotemporal analysis with color coded scale (red), structural (orange), spatial (blue), and temporal (green) aspects of information diffusion.

The construction of information flow among the spatiotemporal clusters are divided into two branches. One highlights source and sink clusters, corresponding to the upper middle box in Figure 3.1, and the other emphasizes information flow, referring to the lower middle box. Here I define the clusters with higher outgoing flows than incoming flows as information source, otherwise as information sink. Then the spatial distribution of the source and sink clusters can be visualized on the world map. Also, I identify the clusters with the largest outflow-inflow differences and consider them as major source and major sink. For all clusters shaped at a time step, the diffusion networks formed by the interactions among their containing retweets are constructed, showing the structural features of the information diffusion of these clusters.

The source-to-sink information flow provides quantitative and geographical description of the information diffusion. At each time step, a matrix showing the flow frequencies of each pair of the clusters is built based on the stream clustering results. Using this matrix, information flow can be presented either in network topology using a heatmap, or in geographical space as a flow map. Heatmap of information flow matrix visually emphasizes flow frequencies among the clusters in the diffusion network, and flow map demonstrates the volume and direction of the flow, and the location of its origin and destination. Besides, based on the information flow matrices at all time steps, a similarity measure is employed to inspect the change of information flow trends in time-series. Further discussion is performed upon the parameter adjustment in stream clustering, information source and sink, and information flow from the scale, structural, spatial, and temporal aspects, with each element color coded in Figure 3.1.

3.3.1 Stream Clustering

Considering the arbitrary spatial distribution of the retweets in both events, GeoDenStream is used to identify the retweets that are close in space and time. Using this method, retweets in the dataset are treated as points distributed in the geographic space, and hence clusters are formed as aggregated points of high density surrounded by those of low density. A major drawback of this clustering method is its heavy reliance on pre-defined parameters, which influence and even determine to a great extent the clustering results. Therefore in this study, great efforts were made on parameter adjustment.

Two key parameters are targeted: a spatial parameter *epsilon* and a temporal parameter time for pruning (*tp*), considering their significant and integrated effect on the clustering results such as cluster size. *Epsilon* refers to the size of the clustering neighborhood, and points within this radius form into one cluster. *Tp* is the time interval for pruning potential- and outlier-clusters. Since the retweets in the datasets scatter all over the world, with latitude and longitude as geographic reference, the *epsilon* parameter is tested from 1 to 10 degrees. As for *tp*, various pruning strategies are applied. First following the default strategy which is by a percentage of point count, I attempt 0.01%, 0.1%, 1% of the total point count in a dataset as the *tp* values. Second, for each tweet I collect all its retweets' timestamps within 24 hours, found the minimum, median, and 75% quantile of the time lags, and average these statistics by day. Then two pruning strategies can be built upon these values: one directly uses these daily average values dynamically as time evolves; and the other uses the global average of these statistics in the whole timeframe (i.e., 84 days of Zika dataset and 120 days of Ebola dataset). The first strategy is considered

point-based, and the last two strategies are time-based. Using combinations of *epsilon* and *tp*, different clustering results are obtained at each time step. To evaluate these results, cluster counts and their visualization on maps are employed.

3.3.2 Information Flow

At each time step, clusters are formed and hence the retweeting activities among these clusters can be identified. Here I only consider the one-hop connection, meaning for each flow, I find its immediate origin and destination disregarding the whole sequence. In practice, this is achieved by detecting the ID of the source tweet embedded in a retweet record. Based on the counts of the retweeting links within each cluster and across different clusters, quantitative measures including information flow matrices and their similarities are applied to support further interpretation of the information flow.

Information flow matrices are a set of squared matrices, with each recording the daily information flow frequencies within and across clusters. At one time step, assume I have n clusters, then an n by n matrix showing directed flow frequencies is constructed, as shown in Table 3.2. In Table 3.2, the diagonal values represent numbers of retweets generated and propagated in the same cluster; and the off-diagonal values count the retweeting links among different clusters.

Table 3.2. A matrix showing the frequencies of directed retweeting links. $F_{i,j}$ denotes the number of retweets in Cluster i and originated from Cluster j.

		Source				
		Cluster 1	……	Cluster j	……	Cluster n
Sink	Cluster 1	$F_{1,1}$		$F_{1,j}$		$F_{1,n}$
	……					
	Cluster i	$F_{i,1}$		$F_{i,j}$		$F_{i,n}$
	……					
	Cluster n	$F_{n,1}$		$F_{n,j}$		$F_{n,n}$

A comparison of the volume and direction of the information flow at different time steps is achieved by a similarity measure (i.e. cosine similarity) that is performed on pairs of flow matrices. Cosine similarity between two vectors (or two documents in the vector space) calculates the cosine of the angle between them. This metric can be interpreted as a normalized comparison between two documents, since it only considers the angle between them without regard to magnitude (Han et al., 2011). To calculate the cosine similarity between two information flow matrices, a prior step is to reshape the two matrices to one-column vectors of the same dimension. For the converted vectors A and B, equation $\cos\theta = \frac{A \bullet B}{||A||\,||B||}$ is applied, where • indicates the vector dot product and ||A|| is the length of vector A.

Appling the above equation, I acquire two types of cosine similarity—consecutive cosine similarity and pairwise cosine similarity. Consecutive cosine similarity is the cos(θ) value of matrices of two adjacent days, which shapes a one-dimensional curve along the timeline. It reveals days with drastic change regarding information flow patterns. Pairwise cosine similarity calculates the cos(θ) of each pair of matrices within the timeframe, which forms a two-dimensional symmetric matrix with each cell containing the similarity measure of its corresponding cluster pair.

3.3.3 Network Properties

In Twitter, information traverses from its originators to receivers, and further diffuses to more audiences. With the development of this process, networks are naturally structured. Hence to understand the pattern of information diffusion, a grasp of the embedded networks' properties, such as general connectivity and power of spreading information to others, is necessary. Therefore, basic network attributes including node and edge counts, network density; and node-associated attributes including degree centrality, closeness centrality, and eigenvector centrality are investigated.

Network density is the ratio of the number of edges in the network over total number of possible edges between all pairs of nodes. It indicates how well a network is connected. Centrality metrics focus on the individual nodes in a network. The more a node connects to others, the greater its centrality is in the network. Degree centrality is defined as the number of edges that a node has. Closeness centrality is the average length of the shortest path between the node and all other nodes in the network. It is "a measure of how long it will take information to spread from a given node to others in the network" (Yin et

al., 2006). Eigenvector centrality is closely related to the influence measure of a node in a network, which is a potential drive of information diffusion. It is based on the principle that the influence of a node depends on the influence of its neighbors (Borgatti, 2005).

3.4 Results

3.4.1 Results for the Zika Outbreak

3.4.1.1 Clustering Results

Applying GeoDenStream, a set of potential values for *epsilon* and *tp* parameters were tested and an optimized combination was determined. Using the selected combination (i.e., *epsilon* = 3 and *tp* = Median), clusters were generated at each time step, and their spatial distributions on the 53rd day, the day with most clusters, are demonstrated on a world map (Figure 3.2 (a)). This map supports the selection of *epsilon*, since it is able to differentiate groups of points in large countries, and aggregate cross-country points covering reasonable sized area, regardless of the shape. By inspecting cluster maps in time sequence, I assured that the cluster IDs of points at the same location are consistent through time. This is crucial for any valid analyses built on this clustering result.

For each cluster, its inflow and outflow frequencies were summed up respectively, and their differences were calculated, based on which clusters were divided into two types: source clusters and sink clusters. In a source cluster, outflow exceeds inflow; and in a sink cluster, its major flow comes from the outside. Figure 3.2 (b) shows the spatial distribution of the source and sink clusters. It suggests that in the discussion about Zika, Twitter users tend to retweet others' information rather than initiating it, except for scattered regions in US, South America, Western Europe, West Africa, Middle East, and Asia.

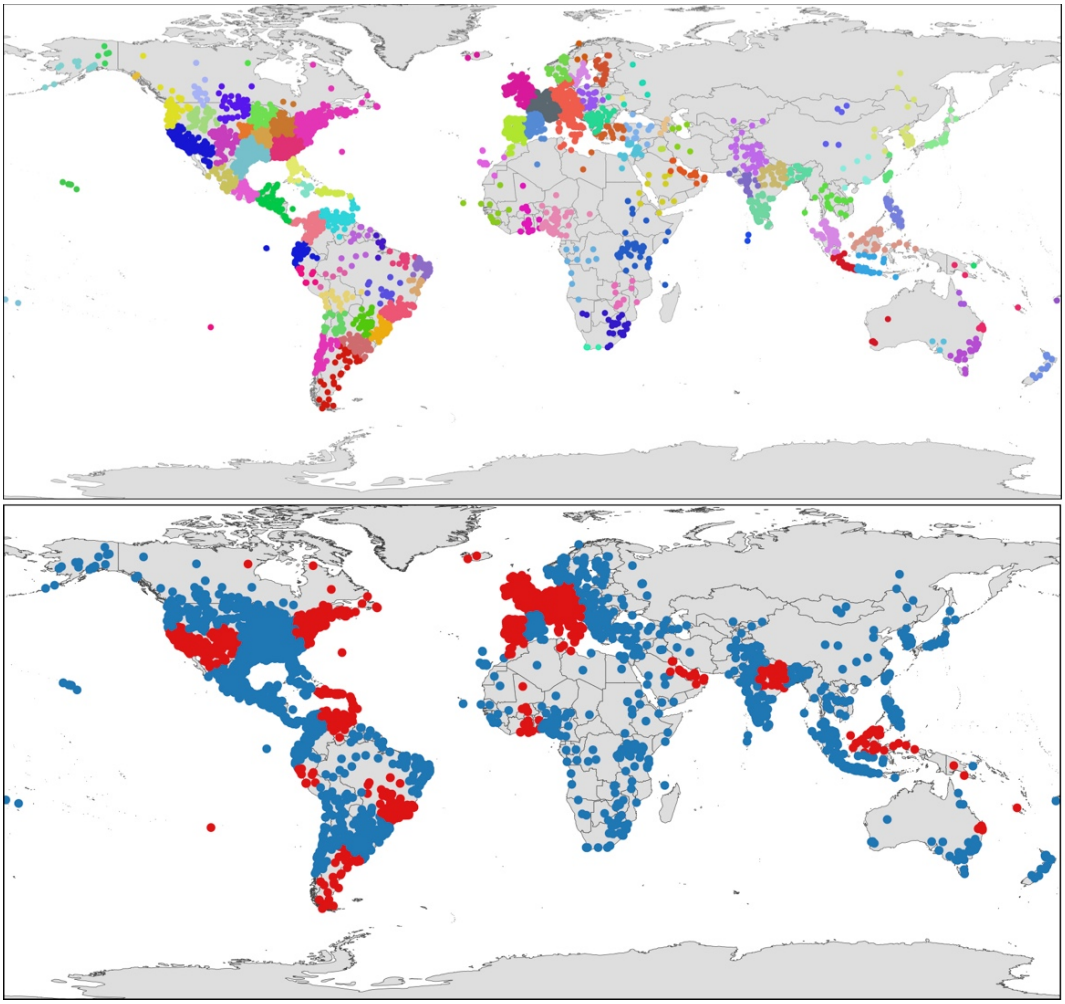

Figure 3.2. Spatial distribution of (a) clusters resulted from GeoDenStream and (b) source (in red) and sink (in blue) clusters overlapped with the world map on the 53rd day in Zika case, when *epsilon* = 3 and *tp* = Median.

Figure 3.3 shows the trend of cluster development over time, along with corresponding tweet counts on each day. Clusters' formation starts from a small number

around 35, then climbs up to about 90 around the mid-stage, then stays around 90 for the rest of the time. Despite several drops in the cluster count curve, its shape generally keeps an upward tendency then levels off. As for tweet count, it stays low and stable for about 35 days, then starts to increases slowly; after a few sharp ups and downs for about 25 days, it gradually declines.

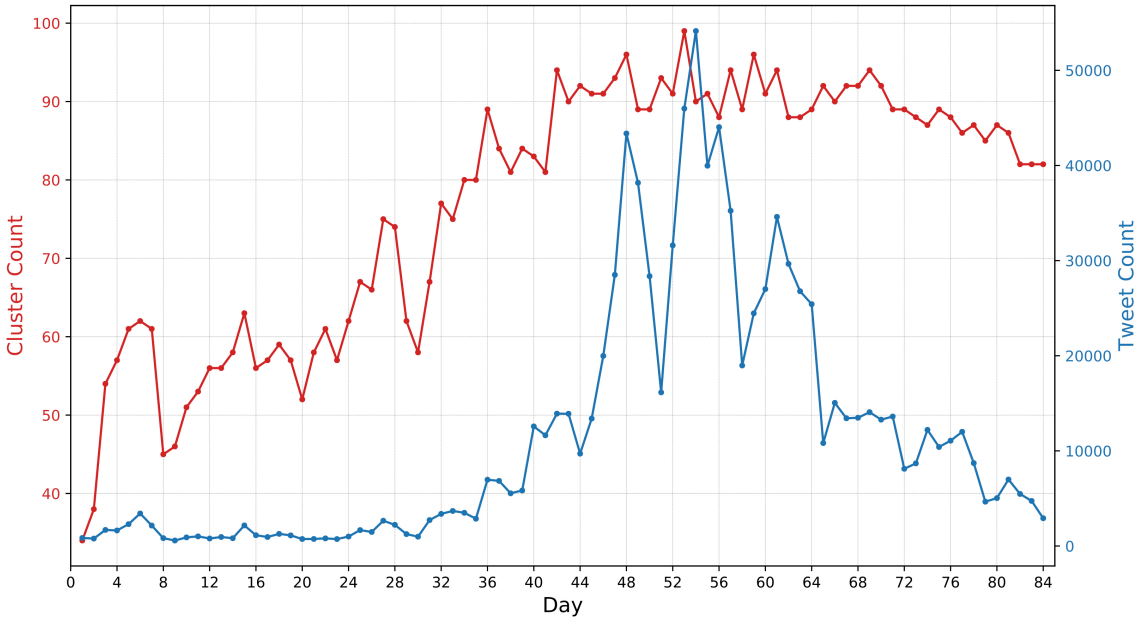

Figure 3.3. Cluster counts and tweet counts when *epsilon* = 3 and *tp* = Median in Zika case.

3.4.1.2 Information Flow Results

Based on the information flow frequencies within each cluster and across different clusters, a series of flow matrices were generated along the time axis. These matrices were

represented by heatmaps showing the numerical values of flow frequency, and by flow maps displaying the spatial distribution of the flows regarding origin, destination, direction, and volume. Flow matrix of the first day is picked as an example to demonstrate its heatmap and flow map (Figure 3.4).

In Figure 3.4 (a) the flow frequencies are normalized to the range of [0, 1] for time-series comparison. The normalization is performed by using the equation $F_i^{norm} = \frac{F_i - \min(F)}{\max(F) - \min(F)}$, where F_i denotes a record of frequency, F_i^{norm} is its normalized value, and $F = \{F_1, F_2, \ldots, F_n\}$. Each cell in Figure 3.4 (a) shows the normalized frequency of the flow from its source in the column and sink in the row. Cells in darker blue reflect higher flow frequencies. From the heatmap it is inferred that large volume of flows mostly occurs within the same clusters; tweets in some clusters such as cluster 30 get more diffused than others, and cluster 33 and 34 have higher tendency of retweeting other clusters. The flow map in Figure 3.4 (b) mirrors the heatmap, adding geographic information including cluster location and direction of each flow. In Figure 3.4 (b), for the best visual experience, only the highest 75% and no more than 30 flow arcs are displayed. The arcs connecting different clusters are drawn in the counter-clockwise direction from the source cluster to sink cluster, and flow within a cluster is represented by the self-directed arc. Cluster IDs are labeled as numbers on the map, and the color bar indicates flow frequency. Production of all flow maps in this chapter follows the same rules.

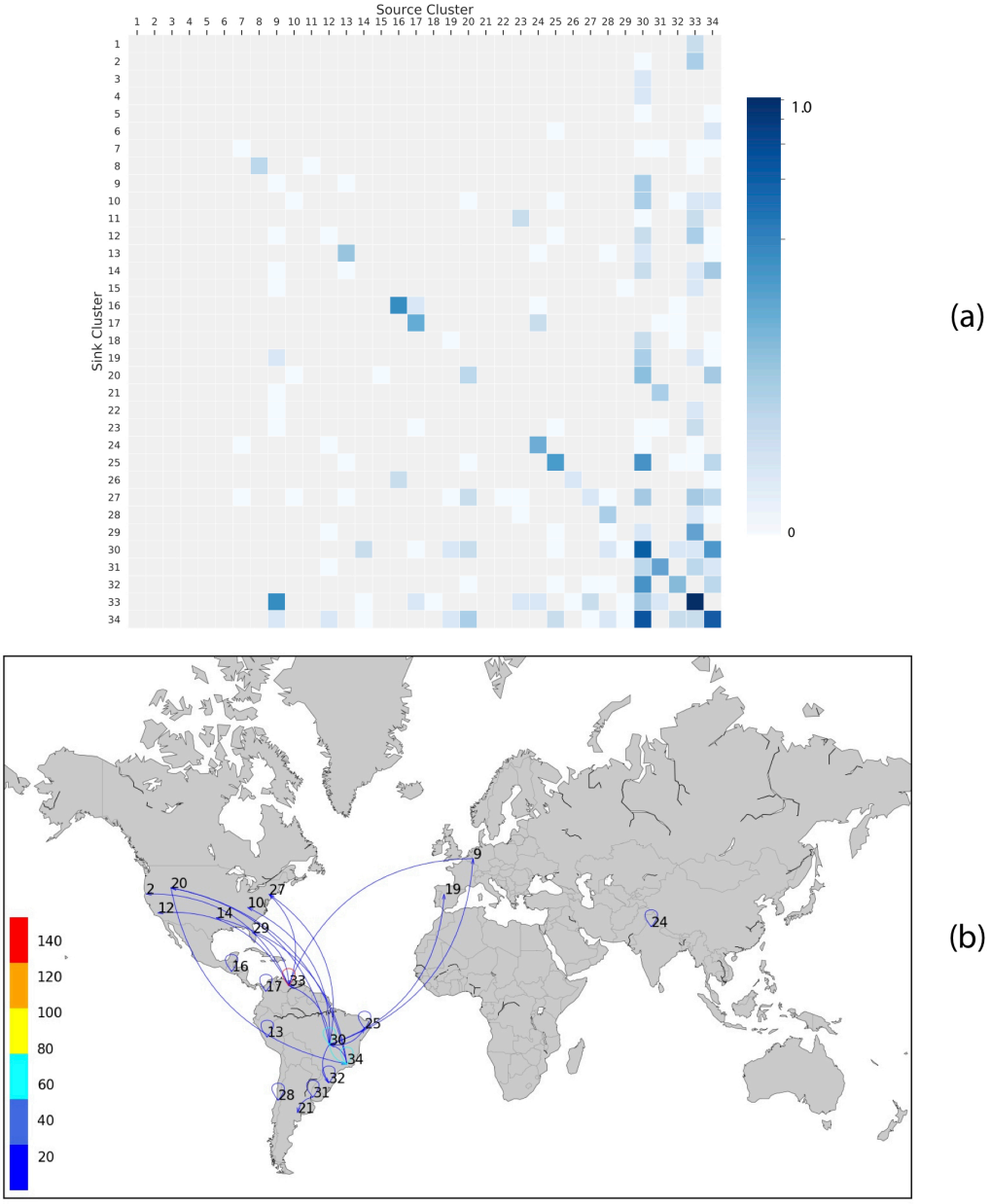

Figure 3.4. Information flow matrix of the first day of captured Zika discussion represented by (a) heatmap and (b) flow map, where arcs are drawn in a counter-clockwise direction between different clusters, and self arcs represent flow within a cluster.

3.4.1.3 Similarity Results

Based on the daily information flow matrices, two types of cosine similarity—consecutive cosine similarity and pairwise cosine similarity—were applied to gauge the change of flow patterns. Consecutive cosine similarity reflects the change of information flow patterns between every pair of adjacent days. In Figure 3.5, the consecutive cosine similarity curve fluctuates greatly for about the first third of the time, then becomes stable for about 40 days before it goes up and down again.

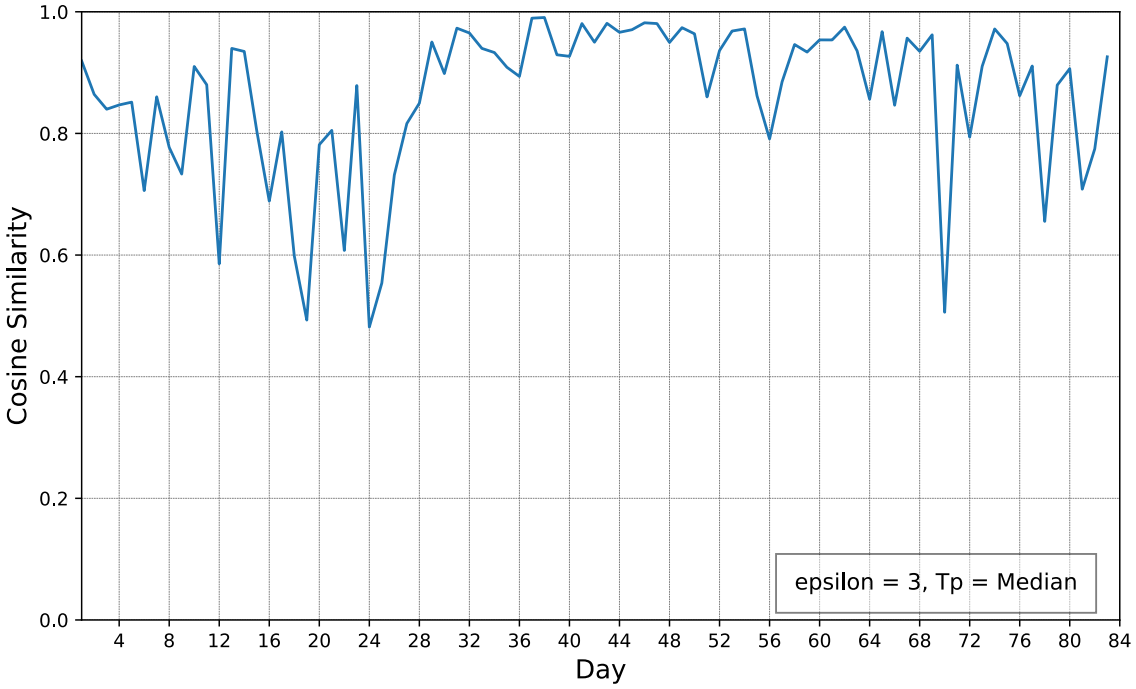

Figure 3.5. Zika consecutive cosine similarity with *epsilon* = 3 and *tp* = Median.

For every pair of days in the studied timeframe, its pairwise cosine similarity value is placed in a symmetric matrix with rows and columns representing the days. Similar to a flow frequency heatmap, the matrix storing the pairwise cosine similarity values is also visualized by a heatmap, shown in Figure 3.6 (a). In the heatmap in Figure 3.6 (a), value in a cell $S_{i,j}$ denotes the cosine similarity between the flow matrices of the i^{th} and j^{th} days. Then a dendrogram, a tree diagram that illustrates the arrangement of the days produced by hierarchical clustering (Everitt and Skrondal, 2010), is created based on the heatmap. Figure 3.6 (b) shows this grouping result of the heatmap, and a tree structure recording the whole grouping process is displayed in Figure 3.6 (c). From the dendrogram I try to identify a certain number of clusters. This is achieved by 'cutting' the tree structure at an appropriate level, and in practice we typically target a huge jump in distance that produces desired group number. For practical concern, I decided to have six groups, and the resulting groups of days are identified by different colors of the tree branches below the cut in Figure 3.6 (c).

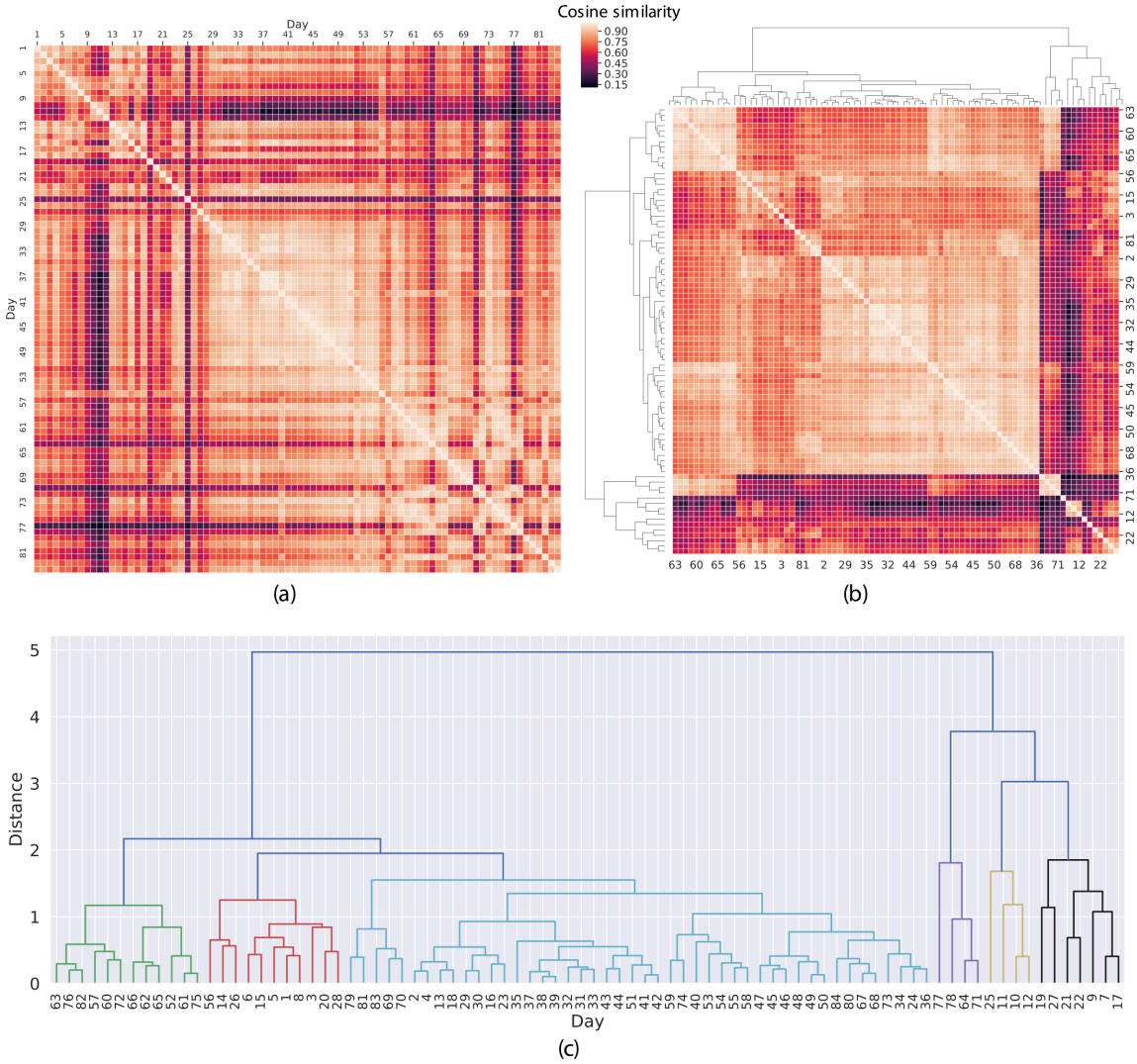

Figure 3.6. Zika pairwise cosine similarity represented by (a) heatmap, (b) dendrogram, and (c) tree structure of hierarchical clustering.

3.4.2 Results for the Ebola Outbreak

3.4.2.1 Clustering Results

Applying the same combination of *epsilon* and *tp* (i.e., *epsilon* = 3 and *tp* = Median), clusters were produced for the Ebola case as well. Figure 3.7 (a) demonstrates a world map showing the spatial distribution of the clusters on the day with most clusters (55^{th} day), where clusters are distinguishable in reasonable sizes, regardless of their shape. As for the distribution of source and sink clusters in Figure 3.7 (b), I learnt that similar to what's found in Zika case, participants in Ebola discussion are also more likely to retweet others' information than to originate it, except for some scattered regions over the world, for example clusters in the US, Western Europe, and West Africa.

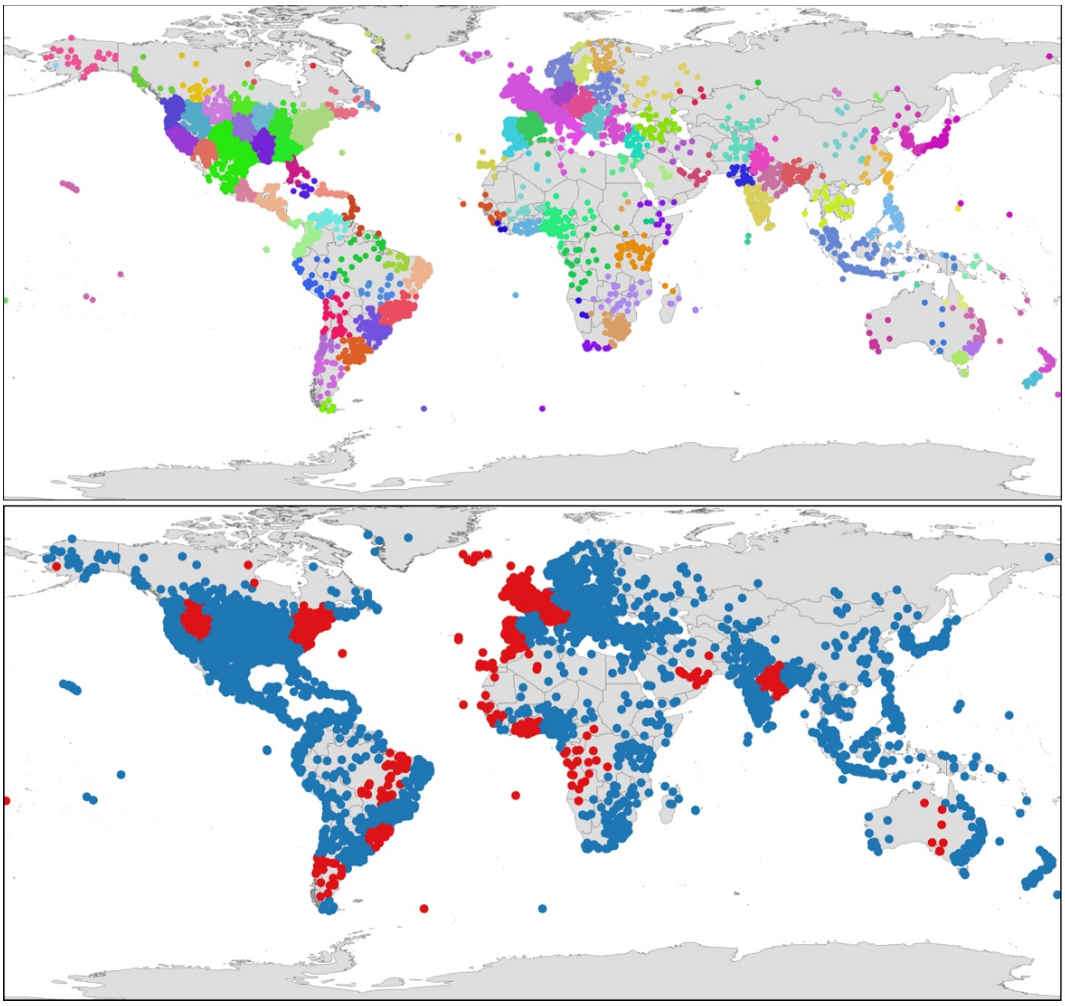

Figure 3.7. Spatial distribution of (a) clusters resulted from GeoDenStream and (b) source (in red) and sink (in blue) clusters overlapped with the world map on the 55th day in Ebola case, when *epsilon* = 3 and *tp* = Median.

Figure 3.8 shows the daily cluster counts and tweet counts of Ebola case. It indicates that different from the former case, clusters start to form quickly from the beginning, then keeps fluctuating around 100 with several spikes. Tweet count stays low

before its first sharp rise on the 41st day, then it grows abruptly to the peak, followed by a few sharp ups and downs. Starting from the 66th day, it declines steadily and levels off.

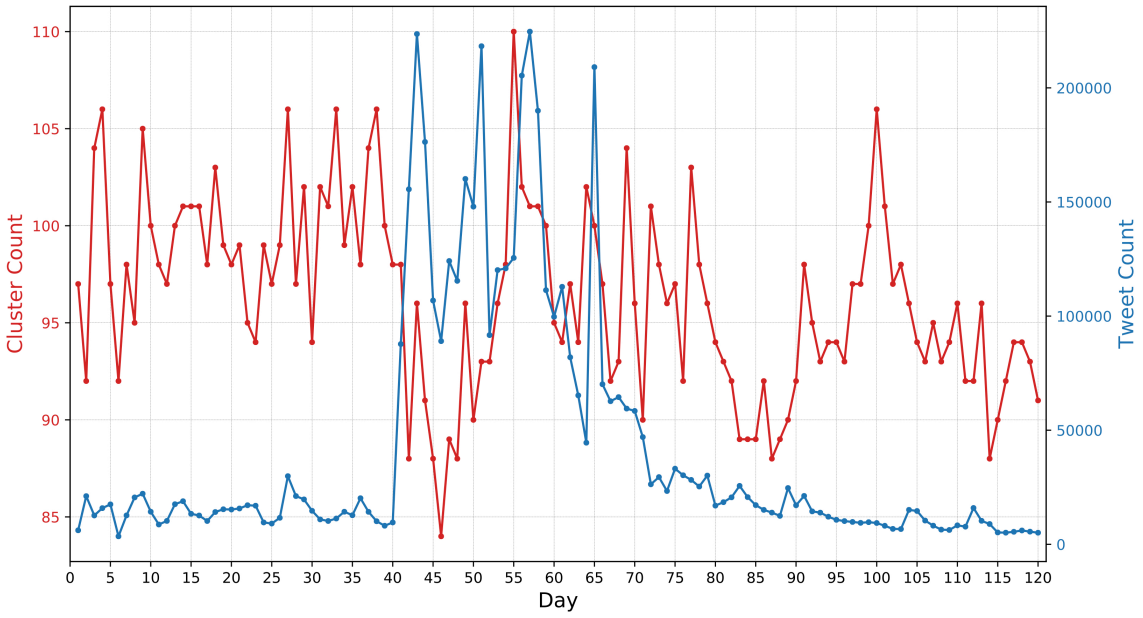

Figure 3.8. Cluster counts and tweet counts when *epsilon* = 3 and *tp* = Median in Ebola case.

3.4.2.2 Information Flow Results

In the captured Ebola discussion, the first day is picked to illustrate its information flow pattern. Using the flow frequencies recorded in the flow matrix, heatmap and flow map were produced and shown in Figure 3.9. In Figure 3.9 (a), each cell shows the normalized frequency of the flow. Darker blue colored cells are mostly found on the primary diagonal line, meaning frequent communication often occurs within the same clusters; and some

clusters such as cluster 94 to 97 interact frequently with each other. Tweets in clusters like cluster 94 and 95 got more retweeted than others, and some clusters are more likely to retweet information from other clusters. Geographic representation of the information flow matrix of the first day is shown in Figure 3.9 (b). On the map, the volume, location, and direction of information flow are clearly visualized, offering a comprehensive explanation of the information flow patterns in the geographical space. The most obvious patterns I detected from this flow map include the active internal retweeing within clusters in West Africa, frequent interaction between clusters in the US and Western Europe, and a major information source cluster in South Asia.

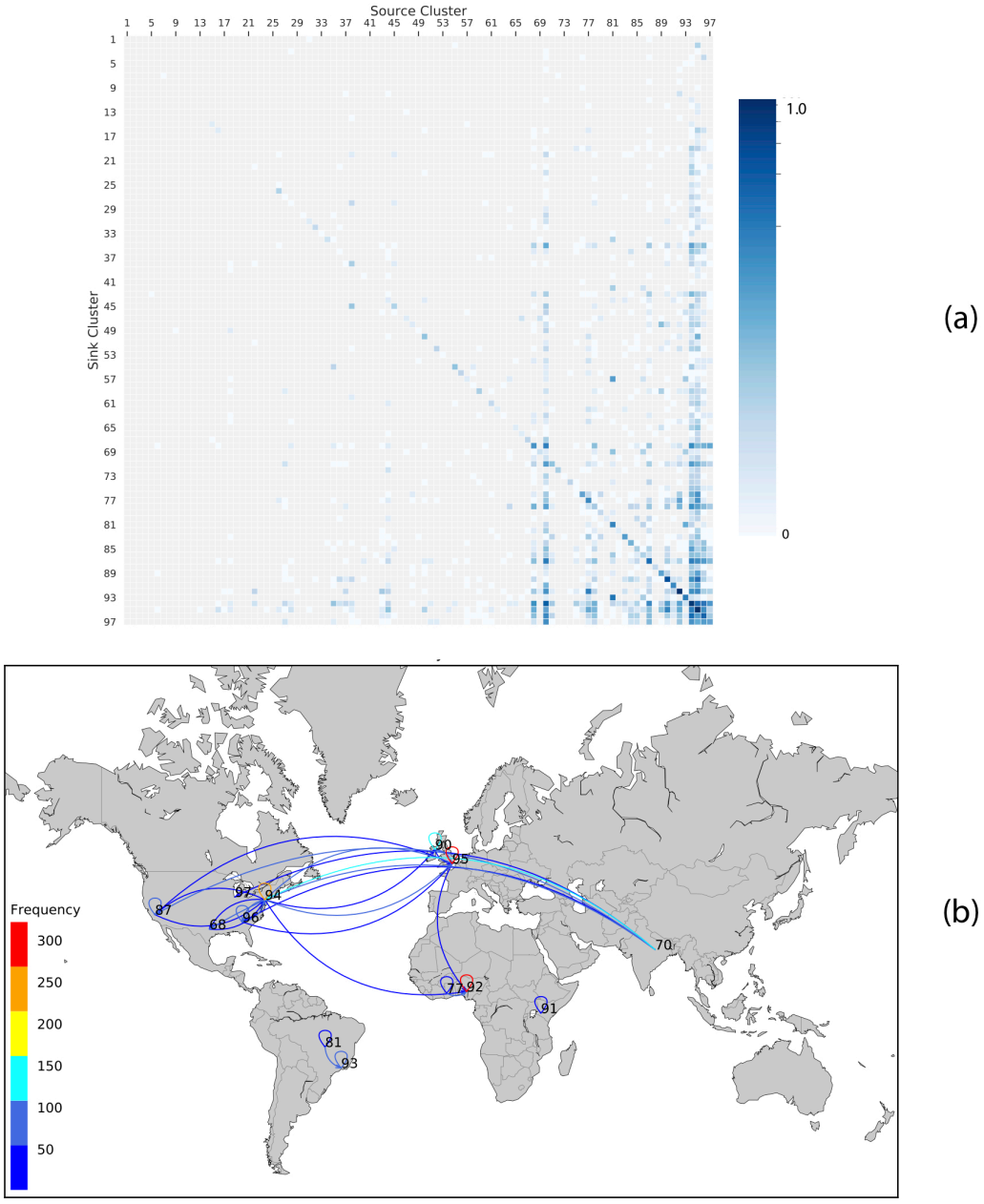

Figure 3.9. Information flow matrix of the first day of captured Ebola discussion represented by (a) heatmap and (b) flow map, where arcs are drawn in a counter-clockwise direction between different clusters, and self arcs represent flow within a cluster.

3.4.2.3 Similarity Results

Consecutive cosine similarity is calculated for all pairs of adjacent days in Ebola case and is plotted in Figure 3.10. Compared with the consecutive cosine similarity of Zika shown in Figure 3.5, this measure of Ebola is generally higher. Another difference lies in the early stage, where consecutive cosine similarity values in Ebola fluctuate modestly at the beginning, while being unstable in Zika. A common pattern of the two cases is that the head and tail of the curves are not as stable as the middle part. Pairwise cosine similarity of the Ebola case is demonstrated as a heatmap in Figure 3.11 (a), a dendrogroam of the heatmap in Figure 3.11 (b), and the structure of hierarchical clustering in Figure 3.11 (c). Same as what was done in Zika, I cut the tree in Figure 3.11 (c) and obtained seven groups of days, colored differently at the branches below the cut.

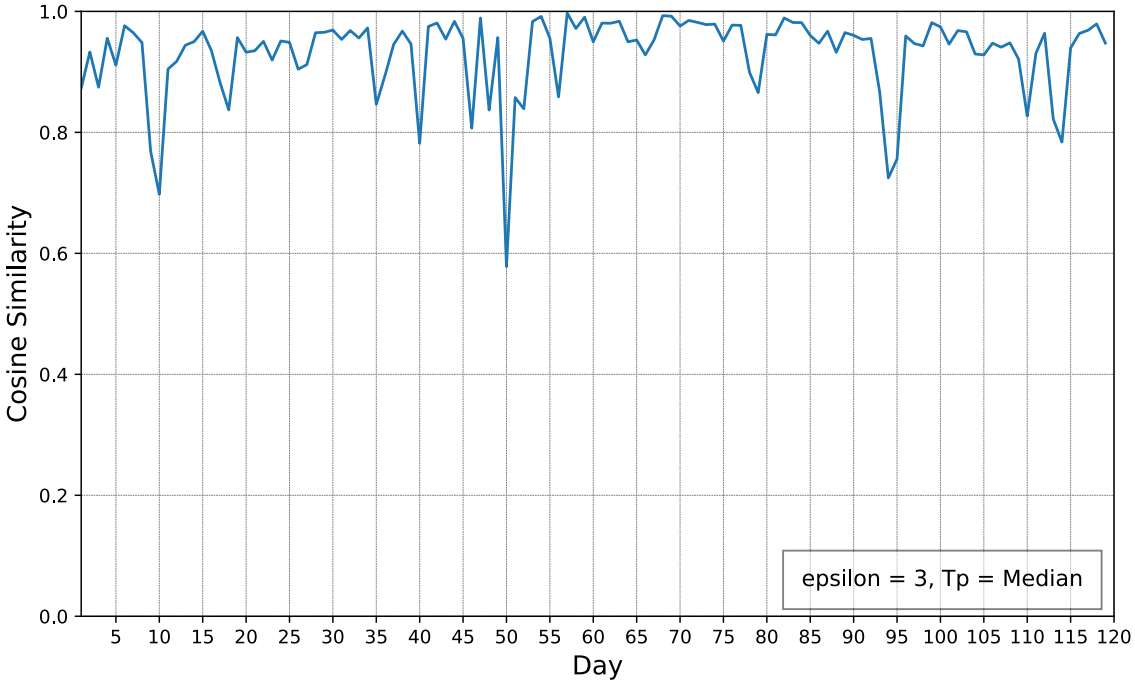

Figure 3.10. Ebola consecutive cosine similarity with *epsilon* = 3 and *tp* = Median.

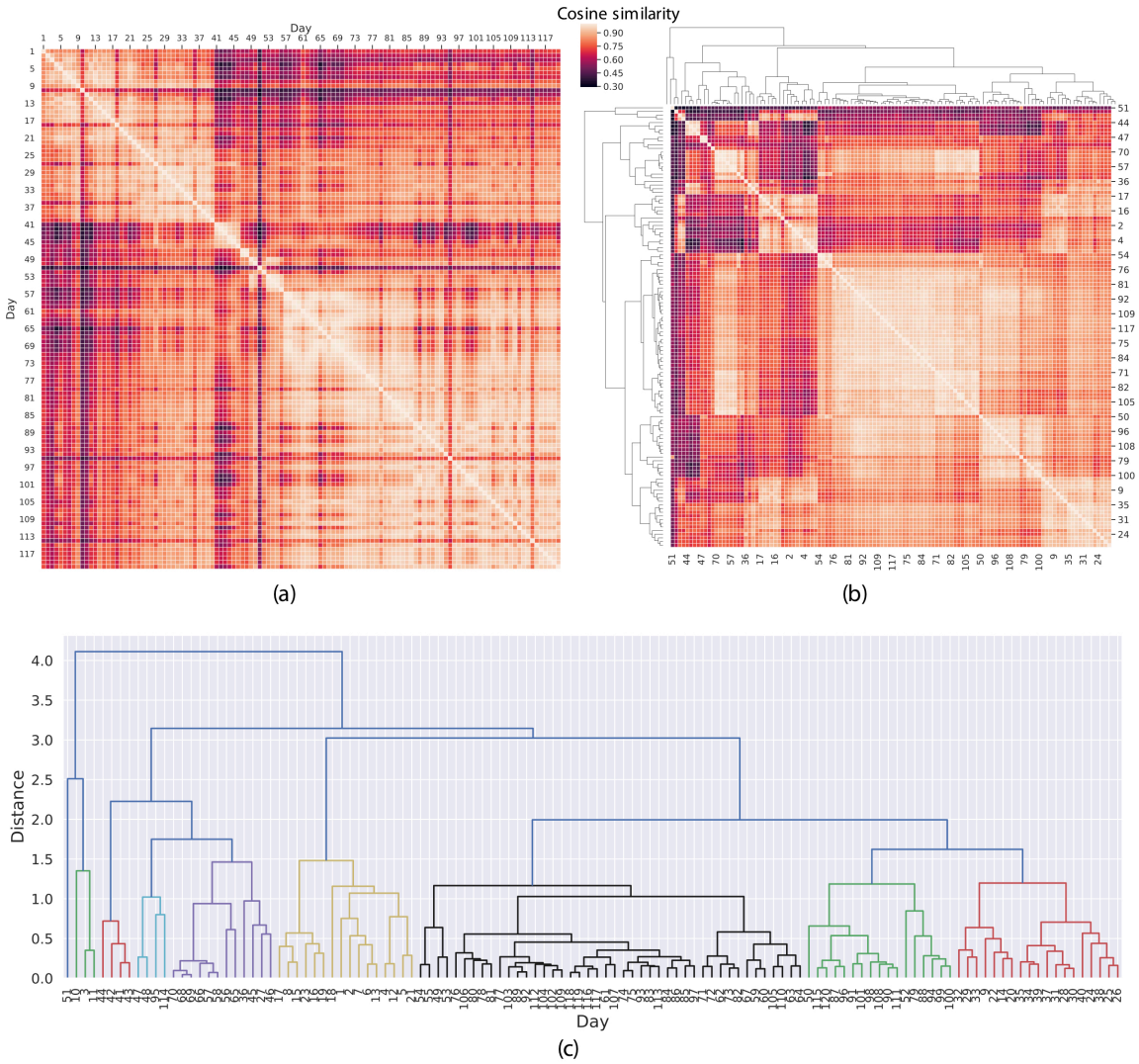

Figure 3.11. Ebola pairwise cosine similarity represented by (a) heatmap, (b) dendrogram, and (c) structure of hierarchical clustering.

3.5 Discussion

In this chapter two widespread epidemics—Zika and Ebola—were chosen for studying the information diffusion patterns within their respective discussion over Twitter via retweeting. These two epidemics occurred in different areas of the world at varied time,

yet both draw global attention. For each event, I separated its related retweets into different clusters and captured the information flow among them at each time step, and calculated the similarity of the flow patterns at different time steps. By further analyzing these results, I expect valuable findings regarding the process of information diffusion in the formed retweeting networks as well as in the geographic environment. The analyses of information diffusion are approached from four perspectives: spatiotemporal scale, structural features, spatial pattern, and temporal evolvement.

3.5.1 Spatiotemporal Scales for Clustering

At the global extent, the influence of spatiotemporal scales on community formation and information transmission among them is complicated. In this study, I gauge geographic scale through the parameter *epsilon* in the stream clustering method, by inspecting cluster counts and consecutive cosine similarity with different *epsilon* settings. Meanwhile, I tackle temporal scale through the parameter *tp* in stream clustering and temporal resolution in data organization.

3.5.1.1 Parameter Setting in the Stream Clustering Method

With large data sets, it is challenging to choose optimal values for the parameters beforehand in the employed stream clustering method. And in practice, we usually find it difficult to integrate real-world meanings into the parameter settings. For example, the parameter *epsilon* is a similarity measure of spatial distance, so a meaningful *epsilon* strongly depends on the distance metric (e.g., degree or meters) and spatial extent (e.g., a local community or a continent) in the data. In the temporal dimension, the parameter time

for pruning (*tp*) requires appropriate assignment considering the temporal scale of data, and targeting at optimal clustering results.

Combinations of *epsilon* and *tp* were tested for both case studies, and curves showing the resulted cluster counts with the parameter combinations are shown in Figure 3.12. For all plots in Figure 3.12-A and 3.12-B, *epsilon* is set to values from {1, 2, 3, …, 10}, meaning the neighborhood radius of developing a cluster is examined from 1 to 10 degrees. From (a) to (c) in Figure 3.12-A and 3.12-B, *tp* values are set based on the count of points, regardless of time. *Tp* used in (d)–(f) are chosen based on the global average of the statistics (i.e., minimum, median, and 75% quantile) of time lags between a tweet and its retweets within 24 hours. And *tp* values in (g)–(i) are dynamic along the time axis, calculated as the daily average of the same statistics.

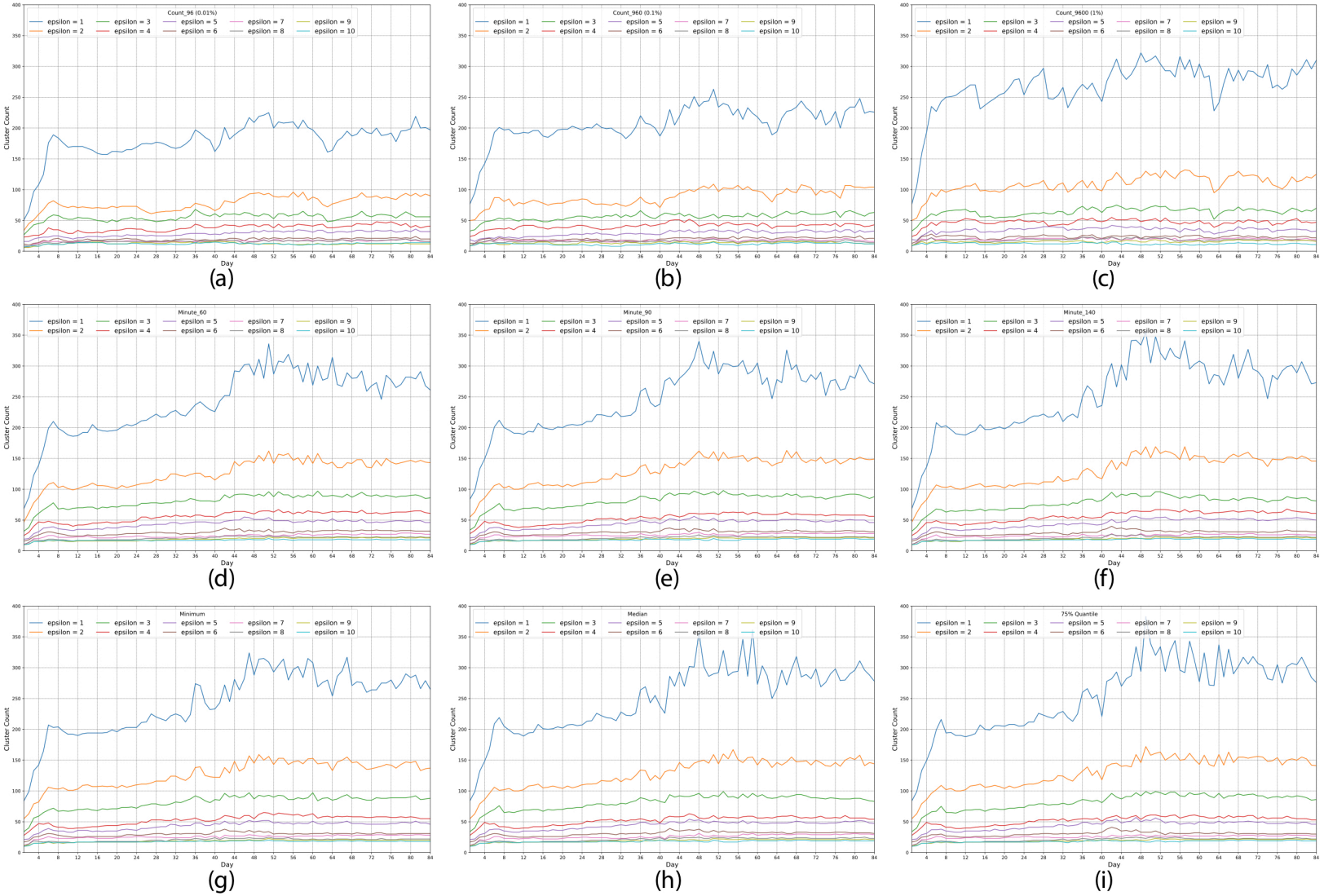
Count_96 (0.01%)
Count_960 (0.1%)
Count_9600 (1%)
Minute_60
Minute_90
Minute_140
Minimum
Median
75% Quantile
epsilon = 1
epsilon = 2
epsilon = 3
epsilon = 4
epsilon = 5
epsilon = 6
epsilon = 7
epsilon = 8
epsilon = 9
epsilon = 10
Cluster Count
Day
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)

Figure 3.12-A. Zika cluster counts with *epsilon* = {1, 2, …, 10} and (a) *tp* = 0.01% of total counts; (b) *tp* = 0.1% of total counts ; (c) *tp* = 1% of total counts; (d) *tp* = globally averaged minimum time lags between a tweet and its retweets within 24 hours (60 minutes); (e) *tp* = globally averaged median time lags between a tweet and its retweets within 24 hours (90 minutes); (f) *tp* = globally averaged 75% quantile of time lags between a tweet and its retweets within 24 hours (140 minutes); (g) *tp* = daily averaged minimum time lags between a tweet and its retweets within 24 hours; (h) *tp* = daily averaged median time lags between a tweet and its retweets within 24 hours; (i) *tp* = daily averaged 75% quantile of time lags between a tweet and its retweets within 24 hours.

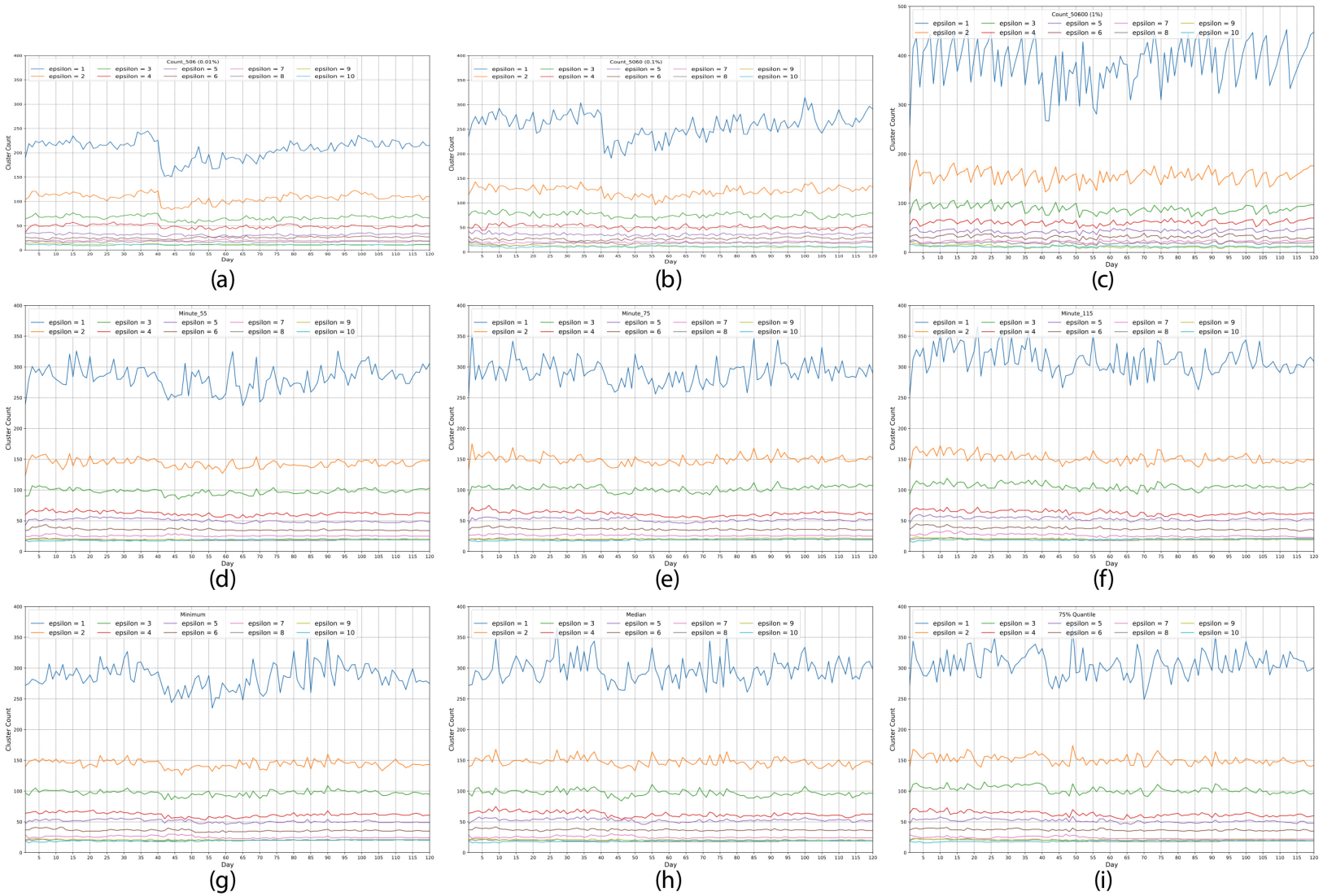
Count_506 (0.01%)
Count_5060 (0.1%)
Count_50600 (1%)
Minute_55
Minute_75
Minute_115
Minimum
Median
75% Quantile
epsilon = 1
epsilon = 2
epsilon = 3
epsilon = 4
epsilon = 5
epsilon = 6
epsilon = 7
epsilon = 8
epsilon = 9
epsilon = 10
Cluster Count
Day
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)

Figure 3.12-B. Ebola cluster counts with *epsilon* = {1, 2, …, 10} and (a) *tp* = 0.01% of total counts; (b) *tp* = 0.1% of total counts; (c) *tp* = 1% of total counts; (d) *tp* = globally averaged minimum time lags between a tweet and its retweets within 24 hours (55 minutes); (e) *tp* = globally averaged median time lags between a tweet and its retweets within 24 hours (75 minutes); (f) *tp* = globally averaged 75% quantile of time lags between a tweet and its retweets within 24 hours (115 minutes); (g) *tp* = daily averaged minimum time lags between a tweet and its retweets within 24 hours; (h) *tp* = daily averaged median time lags between a tweet and its retweets within 24 hours; (i) *tp* = daily averaged 75% quantile of time lags between a tweet and its retweets within 24 hours.

In Figure 3.12-A or 3.12-B, we could find that the patterns of the 10 curves in all plots are identical, with largest cluster numbers when *epsilon* equals 1, and smallest when *epsilon* is 10. Observing across each row, when *epsilon* is fixed and *tp* is set differently by the same strategy—based on count ((a)–(c) or based on time (d)–(i)), the curves differ slightly except for (a)–(c), where cluster count seems to be more sensitive to the selection of *tp*. And the smaller the *epsilon*, the more sensitive cluster count is to *tp*.

Table 3.3 offers a closer look at the cluster counts and their increasing rates with different *tp* strategies for every *epsilon* value, using Zika dataset as an example. It suggests that in most cases, the number of clusters increases with the inclusion of more points for pruning (*tp*), with higher increasing rates than the other two time-based *tp* strategies. The two time-based *tp* adjustment strategies showed similar and relatively low sensitivities of cluster counts to *tp*. In practice, it is usually unrealistic to have global information in the

whole timeframe for calculating the statistics in the second strategy. Therefore, I consider Median in the third strategy as an appropriate setting for *tp*.

Table 3.3. Averaged cluster counts and increasing rates with *epsilon* and *tp* in Zika case.

epsilon	Average Cluster Count			Increasing Rate	
	Count 96	*Count 960*	*Count 9600*	*Count 96~960*	*Count 960~9600*
1	180	207	268	15.00%	29.47%
2	78	88	109	12.82%	23.86%
3	55	56	64	1.82%	14.29%
4	38	41	47	7.89%	14.63%
5	28	29	35	3.57%	20.69%
6	19	20	23	5.26%	15.00%
7	16	18	20	12.50%	11.11%
8	16	16	19	0.00%	18.75%
9	13	14	16	7.69%	14.29%
10	12	12	12	0.00%	0.00%
	Minute 60	*Minute 90*	*Minute 140*	*Minute 60~90*	*Minute 90~140*
1	243	247	253	1.65%	2.43%
2	126	129	130	2.38%	0.78%
3	81	82	78	1.23%	-4.88%
4	56	53	56	-5.36%	5.66%
5	44	44	45	0.00%	2.27%
6	30	30	30	0.00%	0.00%
7	24	26	25	8.33%	-3.85%
8	21	21	21	0.00%	0.00%
9	19	19	19	0.00%	0.00%
10	17	18	18	5.88%	0.00%
	Min.	*Med.*	*75% Qu.*	*Min.~Med.*	*Med.~75% Qu.*
1	244	250	256	2.46%	2.40%
2	127	128	131	0.79%	2.34%
3	81	81	83	0.00%	2.47%
4	53	52	52	-1.89%	0.00%
5	43	44	44	2.33%	0.00%
6	30	31	30	3.33%	-3.23%
7	26	26	26	0.00%	0.00%
8	20	21	21	5.00%	0.00%
9	19	19	19	0.00%	0.00%
10	18	18	18	0.00%	0.00%

Epsilon is mainly decided by the desired cluster number and size in this study. Having too many small clusters is not necessary, and too few is insufficient for reasonable and meaningful analyses. Besides the cluster counts shown in Figure 3.12-A, with additional assistance of map visualization of the clusters (e.g., Figure 3.2 and 3.7) that offers intuitionistic vision of cluster shape and size, I consider the value of 3 suitable for *epsilon*.

Besides *tp*, temporal resolution is another facet of temporal scale. To be specific, in this study data are organized on daily basis, while finer or coarse temporal resolutions could be applied as well. For example, dividing the data streams hourly or weekly. With finer resolution, more details could be captured, yet at the cost of time and computational efficiency. Coarser resolution would erase some details and balance some changes, but would possibly surface the most essential and obvious patterns and phenomenon. Overall, the selection of temporal resolution is essentially empirical depending on the feasibility and our practical demand, and the core principle is to keep enough details without losing an integral understanding.

3.5.1.2 Consecutive Cosine Similarity with Varied Spatiotemporal Scales

In the phase of stream clustering, consecutive cosine similarity reflecting the change of information flow patterns between every two adjacent days was produced with different *epsilon* and *tp* combinations (Figure 3.13). In all plots in Figure 3.13-A and 3.13-B, a common pattern was found that the head and tail of the curves are not as stable as the middle part, meaning that the information flow pattern stays stable at the mid stage while shifts in the beginning and towards the end of the studied time period.

Referring to Figure 3.13, first I focus on the information flow pattern in response to different *tp* settings, reflected by the change of consecutive cosine similarity with the point-based and time-based pruning strategies. This metric seems to show different patterns in the two case studies. In Ebola (Figure 3.13-B), the greater overlap and smaller distinction of curves in (a)–(c) than in (d)–(i) imply that consecutive cosine similarity is less sensitive to *epsilon* when using point count as the pruning strategy, comparing to the two time-based pruning strategies. However, this pattern seems not as obvious in Zika, especially in the second half of time (Figure 3.13-A). In addition, the greater distinction of the curves in early stage (about the first 40 days of Zika and the first 55 days of Ebola) in each plot imply that the flow patterns change greatly in the early time period, then after some time, the patterns of discussion become more stable. Despite the agreement on the overall trend of the curves, a few outliers (i.e., drastic drops) in the plots suggest the integrated role that *epsilon* and *tp* plays in information exchange among the clusters. More specifically, more and sharper drops are found in (a)–(c) than in (d)–(i), suggesting that the point-based pruning strategy incurs greater flow pattern's change in general. This aligns with the higher fluctuation of cluster counts in Figure 3.12 (a)–(c), which indicates larger difference of the daily clustering results. The detection of outliers also assists the selection of *epsilon* and *tp* values: a proper combination should minimize the outliers in the curves.

Next I emphasize the information flow pattern in response to geographic scale, which is indicated by the change of consecutive cosine similarity with *epsilon*. The non-overlapping curves with different *epsilon* values in every plot reveal the influence of geographic scale on this similarity measure, which reflects daily flow pattern's change.

Globally at different geographic scales, the cosine similarities between adjacent days range differently; overall this similarity grows with the decrease of *epsilon* (Figure 3.13). When *epsilon* declines, large clusters are divided into smaller ones, and hence some of the within-cluster flows become across-cluster flows, which will cause the change of information flow patterns at the global level. Such change incurs more details, which sometimes is advantageous, but our overall comprehension could be veiled by the added information. Therefore, geographic scale is influential on the cluster-level information flow, and should be handled carefully.

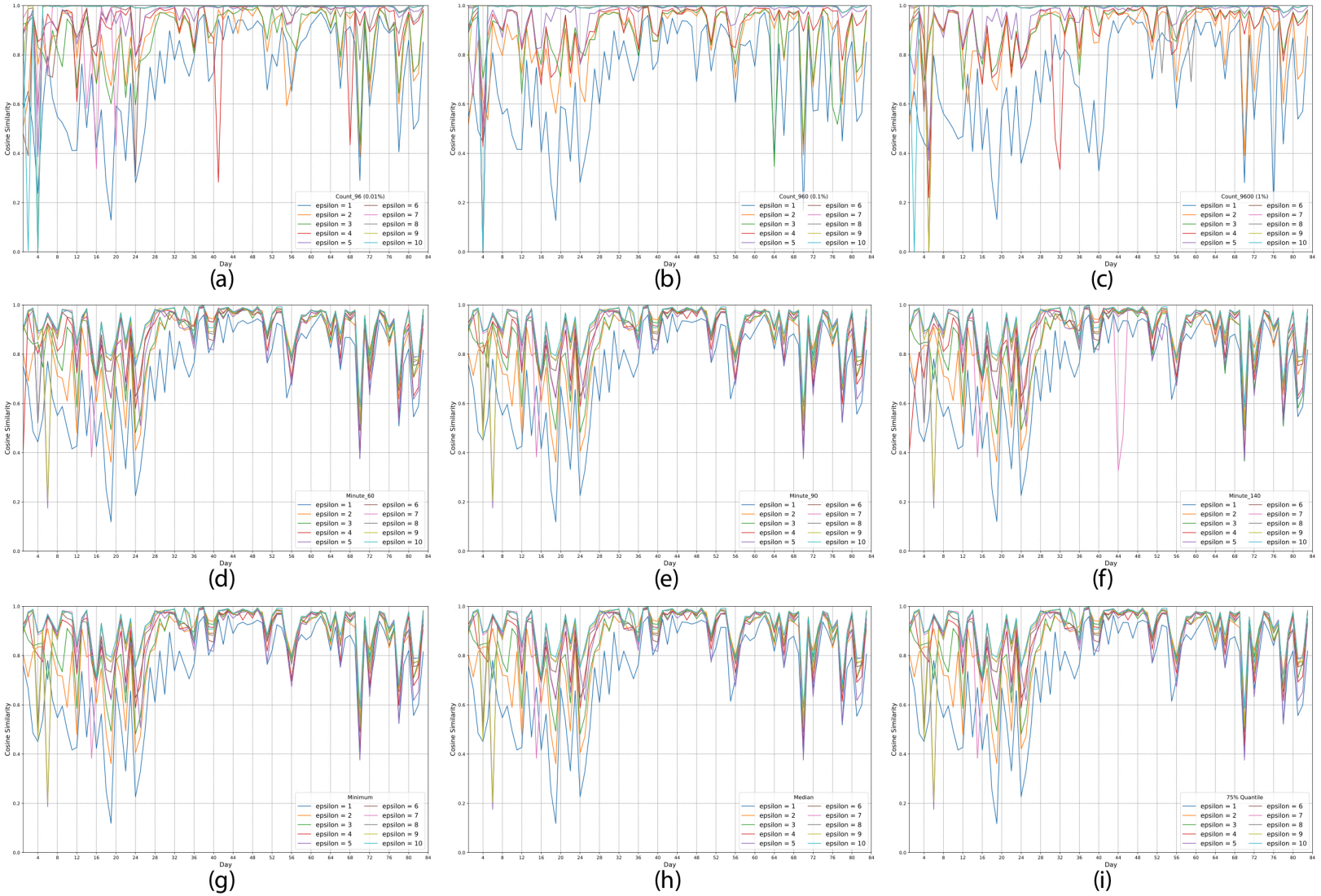
Cosine Similarity
Day
Count_96 (0.01%)
Count_960 (0.1%)
Count_9600 (1%)
Minute_60
Minute_90
Minute_140
Minimum
Median
75% Quantile
epsilon = 1
epsilon = 2
epsilon = 3
epsilon = 4
epsilon = 5
epsilon = 6
epsilon = 7
epsilon = 8
epsilon = 9
epsilon = 10
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)

Figure 3.13-A. Zika consecutive cosine similarity with *epsilon* = {1, 2, …, 10} and (a) *tp* = 0.01% of total counts; (b) *tp* = 0.1% of total counts; (c) *tp* = 1% of total counts; (d) *tp* = globally averaged minimum time lags between a tweet and its retweets within 24 hours (60 minutes); (e) *tp* = globally averaged median time lags between a tweet and its retweets within 24 hours (90 minutes); (f) *tp* = globally averaged 75% quantile of time lags between a tweet and its retweets within 24 hours (140 minutes); (g) *tp* = daily averaged minimum time lags between a tweet and its retweets within 24 hours; (h) *tp* = daily averaged median time lags between a tweet and its retweets within 24 hours; (i) *tp* = daily averaged 75% quantile of time lags between a tweet and its retweets within 24 hours.

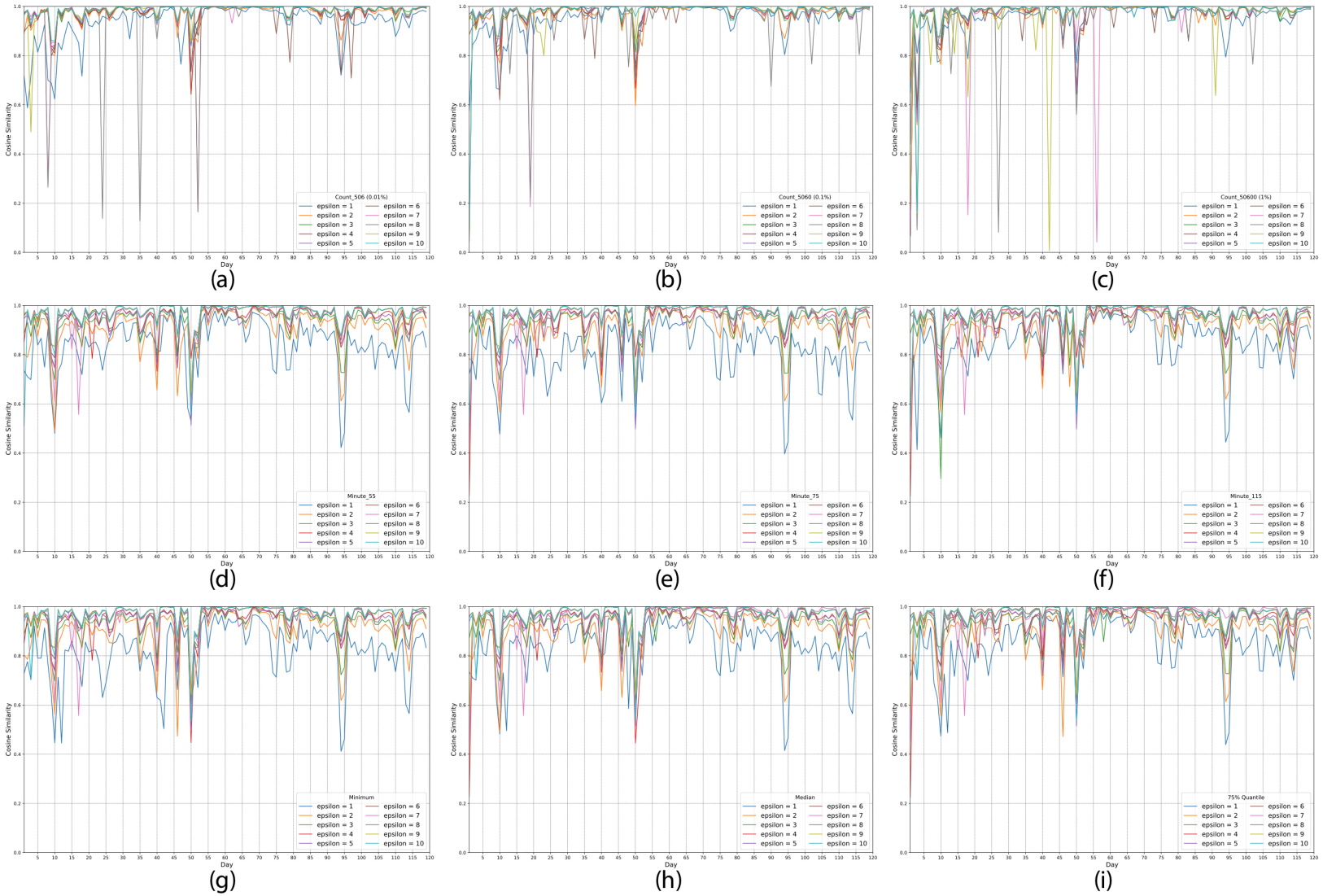
Cosine Similarity
Day
Count_506 (0.01%)
Count_5060 (0.1%)
Count_50600 (1%)
Minute_55
Minute_75
Minute_115
Minimum
Median
75% Quantile
epsilon = 1
epsilon = 2
epsilon = 3
epsilon = 4
epsilon = 5
epsilon = 6
epsilon = 7
epsilon = 8
epsilon = 9
epsilon = 10
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)

Figure 3.13-B. Ebola consecutive cosine similarity with *epsilon* = {1, 2, …, 10} and (a) *tp* = 0.01% of total counts; (b) *tp* = 0.1% of total counts; (c) *tp* = 1% of total counts; (d) *tp* = globally averaged minimum time lags between a tweet and its retweets within 24 hours (55 minutes); (e) *tp* = globally averaged median time lags between a tweet and its retweets within 24 hours (75 minutes); (f) *tp* = globally averaged 75% quantile of time lags between a tweet and its retweets within 24 hours (115 minutes); (g) *tp* = daily averaged minimum time lags between a tweet and its retweets within 24 hours; (h) *tp* = daily averaged median time lags between a tweet and its retweets within 24 hours; (i) *tp* = daily averaged 75% quantile of time lags between a tweet and its retweets within 24 hours.

3.5.2 Structure of Information Diffusion

Understanding the structure of the diffusion network is important for studying information flow pattern, since it lays the foundation for information diffusion activities rooted in the network. Therefore, to explore the structure of the network generated from retweeting, selected network attributes are calculated for each day and plotted in Figure 3.14. Attributes in the upper two rows are measures of the network structure from different perspectives. The lower three rows illustrate centrality measures associated with individual nodes in the network, indicating the averaged nodes' characteristics.

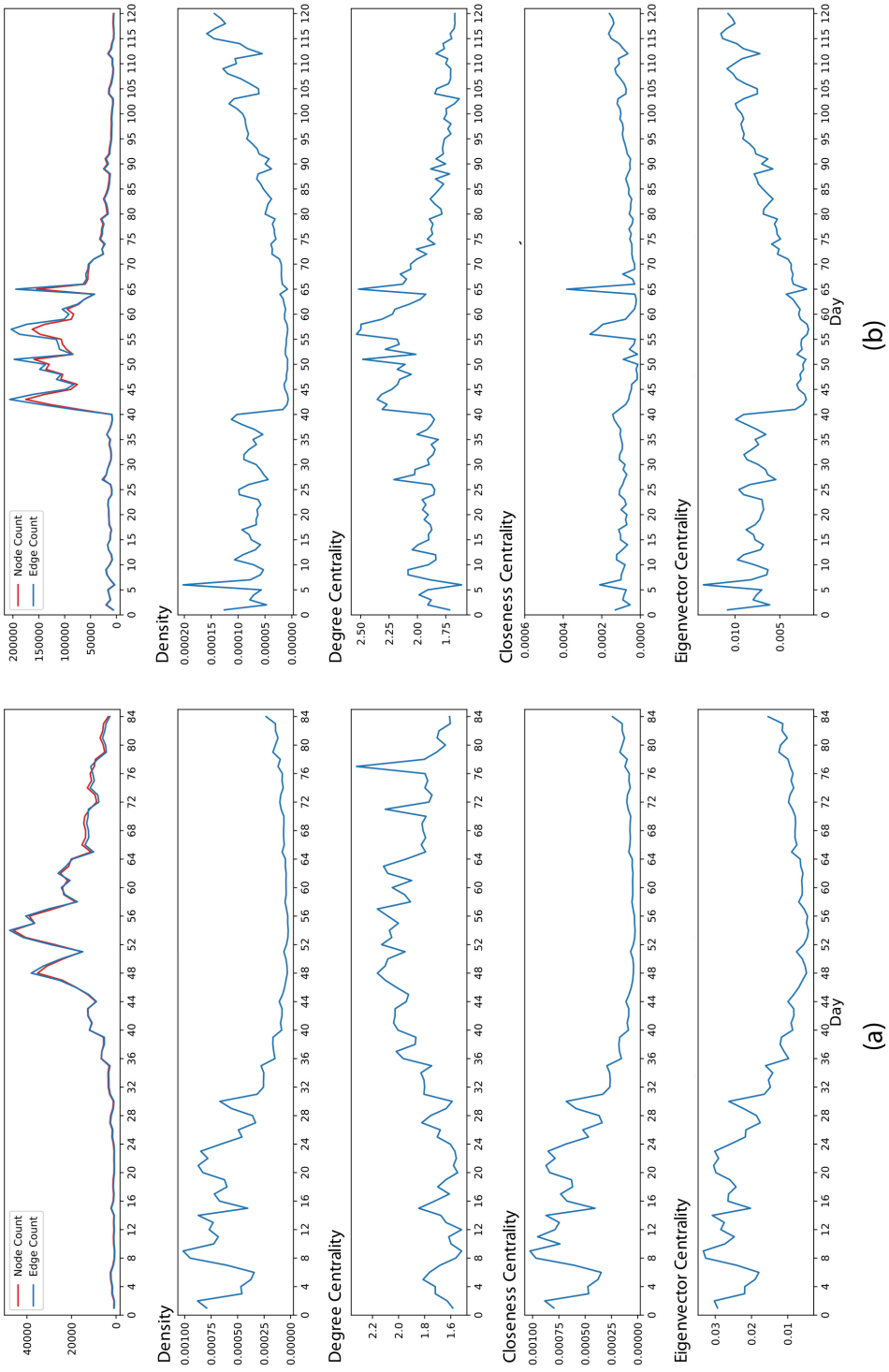

Figure 3.14. Daily network properties of (a) Zika and (b) Ebola.

The network properties graphs reveal some interesting trends. The evolving trends of network density suggest that in the early stage, there are not so many participants involved in the discussion as shown by node count, yet are well connected because of their common interest in the topic. Then with the growing popularity of the topic, more and more Twitter users joined the discussion, but their loose connection lead to the decline of network density. After a while, due to the loss of attention from the general public, the ones stayed in the discussion are actually concerned about the event, and their active communication with each other have prompted the increase of network density. In the later stage of Ebola, since the discussion has subsided for a while, participants in this phase are actually interested in the topic and well connected with each other, which caused the growth of density in this period. However, in the later phase of Zika, this topic has gradually receded but still drawn wide attention, indicated by the large amount of nodes, therefore network density is still low during this period due to their loose connection. If a longer time period was captured in the Zika case, I would expect a similar uprising tendency of density in its later stage. However, this guess needs to be further examined because of the general impression that the impact of Zika include countries with better internet connection (e.g., the US, Brazil, etc.), and the discussion is among the general public; while on the other hand discussion of Ebola remains mostly within "experts".

The trend of degree centrality is very similar to that of node and edge counts. Degree centrality represents the averaged measure of indegree and outdegree, showing a node's connectivity with other nodes. Suggested by the curves, degree centrality rises as

the number of participants increases. At the same time, the fluctuation of degree centrality indicates the participation of new users and the exit of existing users.

Closeness centrality explains the shortest path for a node to diffuse a message to all other nodes in a network. The shorter the path, the higher the closeness centrality. Therefore in some cases, more participants could decentralize the network and reduce the average of this measure. Because of the larger population of participants in Ebola discussion than in Zika, closeness centrality of Ebola network is much smaller. The trend of closeness centrality of Zika is coherent with density, because denser network is usually better connected, causing higher closeness centrality. This overall trend is also found in Ebola, except that since large population of participants joined the network on the 57^{th} and 65^{th} days, connections in the network changed rapidly and possibly compelling information disseminators emerged, resulting in abrupt increase of closeness centrality on these days.

Same as closeness centrality, eigenvector centrality is also coincident with density regarding overall trends. Eigenvector centrality reflects the influence of the nodes on information diffusion. As more participants join in the discussion, the voice of major information spreaders is counteracted by the large volume of mass communication with lower eigenvector centrality, and hence its averaged value would drop. In later phase when general audiences have left, the influence of the ones that remain would be promoted, so the averaged eigenvector centrality of the network would increase. In sum, the network properties I chose are able to help describe the information diffusion patterns.

Nodes in the network could be anchored in the geographical environment, and the edges formed by information flow connecting these nodes could also be visualized

geographically. Aiming at an overall geographical perception of the information diffusion network of the two case studies, the flow frequencies are accumulated at the cluster level and drawn on world maps in Figure 3.15. For the sake of computational efficiency, the first 30 days were counted.

Maps in Figure 3.15 reveal that in the discussion about Zika, the largest cluster-level accumulated information flow happens within the clusters located in Venezuela and Brazil: both at high risk of Zika virus disease. The largest accumulated information flow of Ebola also occurs within clusters, located in the UK and West Africa, with West Africa being the origin and highly infected area of Ebola virus. Also, discussion within the cluster in the northeastern US possesses large volume. As for dominant across-cluster flow, I detected evident difference between the two case studies: Zika is widely discussed among clusters in North and South America, and Ebola is popular among North America, Western Europe, and West Africa. The places with extensive discussions align well with the high-risk countries of the respective disease in real-world situations.

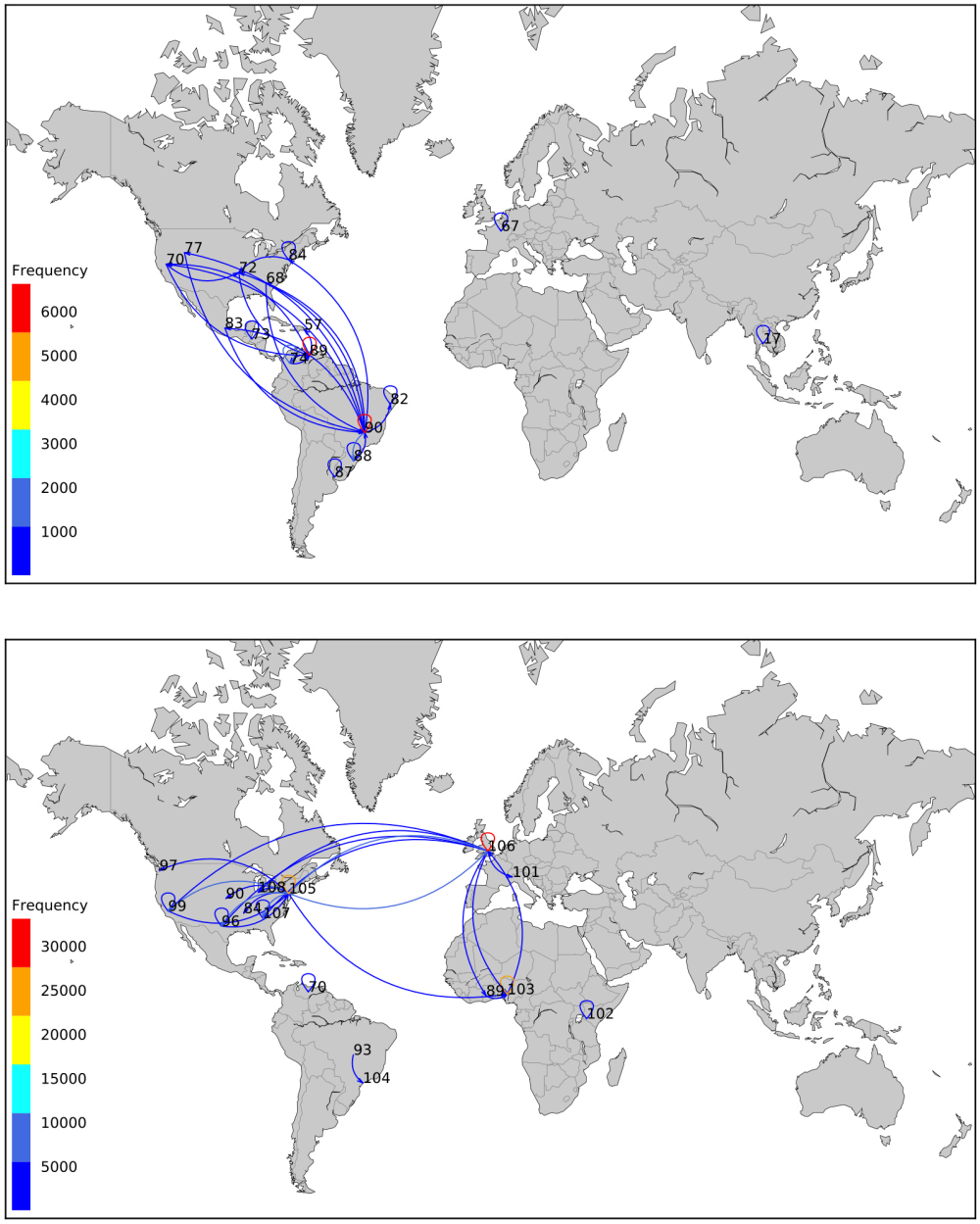

Figure 3.15. Accumulated information flow for the first 30 days of (a) Zika and (b) Ebola.

3.5.3 Spatial Patterns of Information Diffusion

3.5.3.1 Spatial Distribution of Major Source and Sink

From the spatial distribution of source and sink clusters shown in Figure 3.2 (b) and Figure 3.7 (b), it can be inferred that for both events, most of the participants tend to retweet others' information rather than initiating it. Further information could be obtained from Figure 3.16: among all sink clusters, the top four with the largest inflow-outflow differences in both cases are all located at the mid-east part of the US; and among all source clusters, the top four with the largest outflow-inflow differences cover areas in the US, part of Western Europe, and central Brazil in both events. A special major source cluster in Ebola is located in West Africa.

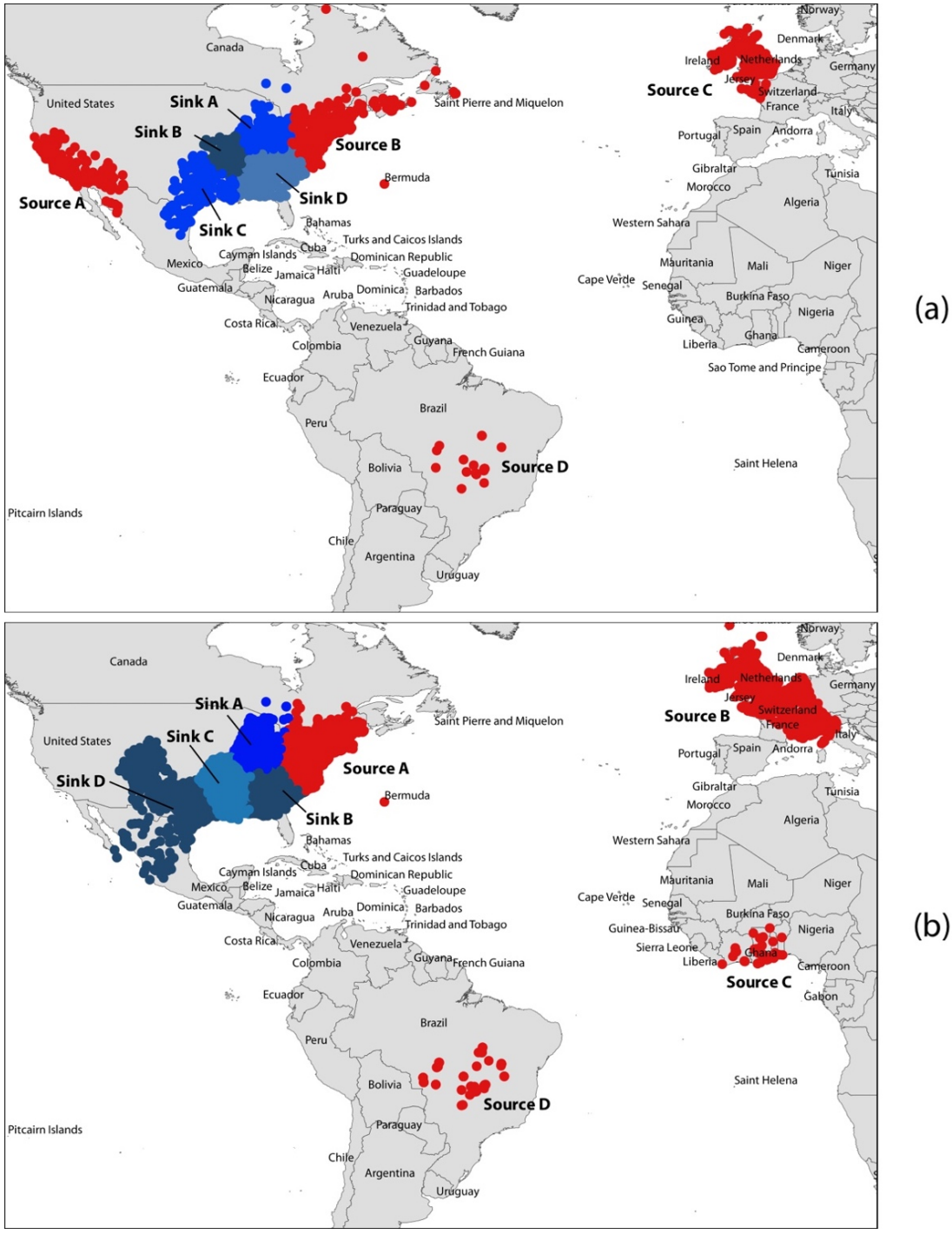

Figure 3.16. Top four source clusters (in red) and sink clusters (in blue) on world map of (a) Zika and (b) Ebola.

From the studied events, we can generalize that participants in the US are the most active ones regarding both generating information and spreading it. It is not surprising since according to PeerReach (2013), the US heads the list of monthly active tweeting users by a wide margin, followed by Japan, Indonesia, UK, and Brazil. The large population of active tweeting users ensures the vigorous publishing and retweeting activities over Twitter in the US. Running down this country list, I found that for the studied cases, users in the UK and Brazil are also active in creating tweets; however users in Japan seem not as interested, and users in Indonesia are more concerned about producing information about the Zika outbreak. From the spatial distribution of major source and sink clusters, I infer that users' spatial distribution and their activeness in Twitter play important roles in the information diffusion process.

Besides Twitter users' distribution and activeness, another important drive of information diffusion is the actual location where the outbreaks take place. A similar conclusion was drawn in Kwon et al. (2015), where it was stated that "transnational information diffusion can be influenced by spatial proximity between the origin nation and other parts of the world." More interestingly, in my study, the event location usually shows higher significance in the early stage, because in early phase people closer to the event are more likely to show instant interest. This statement is supported by the large flow volume from the very beginning in Brazil where Zika prevails, and in West Africa where Ebola breaks out (Figure 3.19 (a) and (e)). When the information spreads more widely at later time, it draws the attention from people at greater spatial extent, thus at this time event location is not as influential as before. The locations of major source and sink clusters

outside the event location such as US and Western Europe support this finding (Figure 3.16).

The influence of geolocation also explains why in both cases Japan seems not as involved as other countries on the list provided by PeerReach (2013) (Figure 3.2 (b) and Figure 3.7 (b)): its long distance to the event locations despite its highly active tweeting users; and why Indonesian users are more active in Zika discussion (Figure 3.2 (b) and Figure 3.7 (b)): its location near the equator, where the tropical climate intensifies the threat from the mosquito-borne Zika virus to this region. This is perceived as geography-driven homophily that similar geographic situation of Indonesia and high-risk areas of Zika such as Brazil has invoked similar level of retweeting activeness.

3.5.3.2 Information Flow Distribution of Major Source and Sink

To discover the spatial distribution of information flow of major source and sink clusters, four clusters were selected as examples, including the major source cluster located in Brazil in Zika event (Source D in Figure 3.16 (a)) and the major source cluster located in West Africa in Ebola event (Source C in Figure 3.16 (b)), and one major sink cluster in the US in both events (Sink A in Figure 3.16 (a) and (b)). The daily outflow maps of the two major source clusters, and daily inflow maps of the two major sink clusters were generated; and among them one day that best represents the overall pattern is picked for visualization in Figure 3.17.

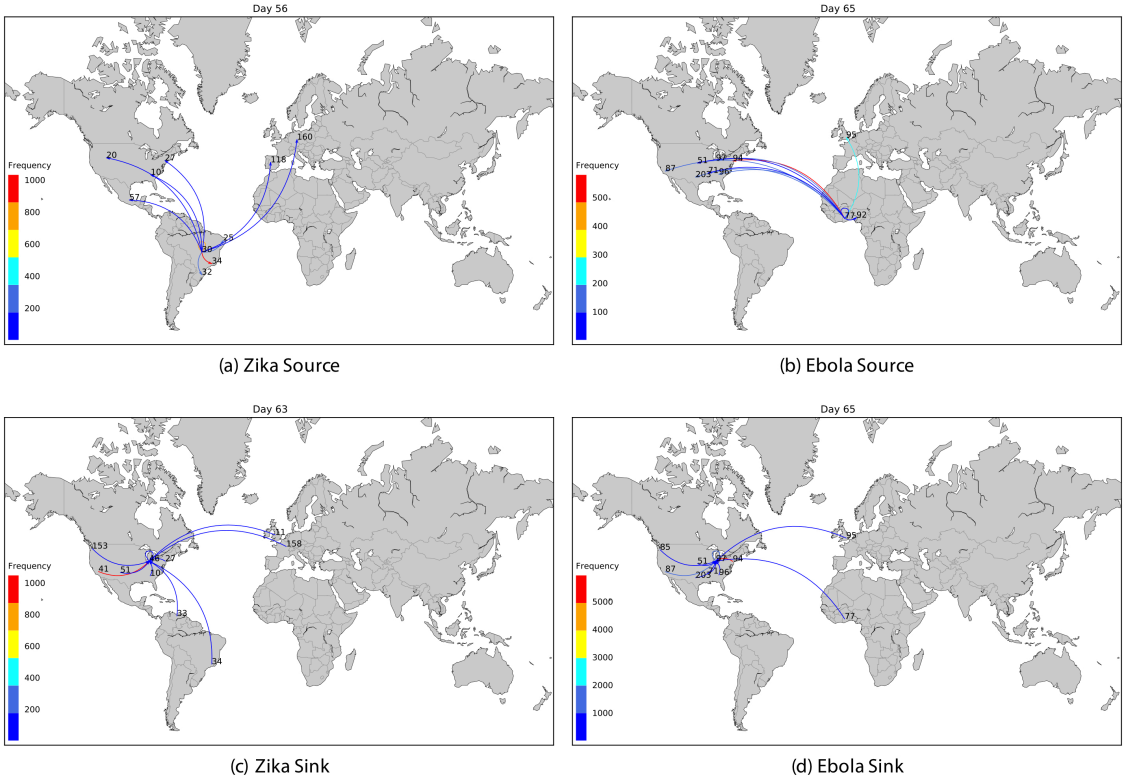

Figure 3.17. Flow maps of the major source cluster in Brazil in Zika case (a) and West Africa in Ebola case (b), and one major sink cluster in the US in Zika case (c) and in Ebola case (d).

By inspecting the daily outflow/inflow maps of the four clusters, I spotted the areas of clusters that most frequently interacted with them. In the Zika discussion, information originated in Brazil mainly diffused to its neighboring clusters in Brazil, the US, and Western Europe (Figure 3.17 (a) as an example); and the major sink in the US primarily received information from American west, Brazil, Venezuela, and Western Europe for most of the time (Figure 3.17 (c) as an example), and from dispersed areas all over the

world in later stage. In the Ebola discussion, information produced in the major source cluster in West Africa mostly reached to the US and Western Europe (Figure 3.17 (b) as an example); and the information spread to the chosen major sink in the US was predominantly from other areas in the US, as well as Western Europe and West Africa (Figure 3.17 (d) as an example). In addition, I noticed that Twitter users in the US and Western Europe actively participated in the discussion of both events, via generating and retweeting the relevant information. And the locations where events actually took place—South America in Zika and West Africa in Ebola—are important birthlands of the respective information.

From the information flow maps produced earlier (Figure 3.4 (b), 3.9 (b), 3.15, and 3.17 as examples), I found that retweeting widely exists within and across clusters, and further summarized three key findings about information flow. First large information flow occurs among active tweeting areas; a typical situation is the communication among major sources; for instance, the frequent flow between northeastern US and the Western Europe in both events. Second is the large flow between event location and other places especially active tweeting areas. In this case the event location normally acts as the information source; for example, flow from Brazil and directed to the US and Western Europe in Zika, and flow from West Africa to the US and Western Europe in Ebola. Third, Tobler's First Law of geography which stresses the effect of physical proximity (Tobler, 1970), in this case translated to the tendency of retweeting from physically close users, does not always hold true. Users in one cluster not only retweet others in the same cluster and in their

adjacent clusters, but also maintain active across-cluster and across-sea communications. Instead, geography-driven homophily has a greater impact.

3.5.4 Temporal Evolvement of Information Diffusion

3.5.4.1 Temporal Evolvement of Information Flow Patterns

The change of information flow patterns over time is inferred from the two types of cosine similarity measures. From the consecutive cosine similarity curves in Figure 3.5 and 3.10, a few drastic changes could be detected. For each event, the day with the most significant change (i.e., differs the most with one day before and after) of consecutive cosine similarity (i.e., the 70th day in Zika and the 50th day in Ebola) was chosen, and the information flow maps of one day before and two days after it were scrutinized. They correspond to the 69th~72nd days (February 18th~21st, 2016; Thursday to Sunday) in Zika, and 49th~52nd days (October 8th~11th, 2014; Wednesday to Saturday) in Ebola. Information flow maps of these days are displayed in Figure 3.18, providing a geospatial perspective of the patterns change.

In Zika case, the dive of consecutive cosine similarity value on the 70th day means that the information flow matrix of the 69th and 70th days are similar, and so are the 71st and 72nd days; however, the 70th and 71st days have different patterns. This is supported by the change of geographical distributions of the information flow in Figure 3.18 (a)–(d), where an obvious distinction between (a)–(b) and (c)–(d) is the shift of the most frequent outflow from northeastern US to southwestern US. Similarity, the most evident discrepancy between (e)–(f) and (g)–(h) lies in the emergence of clusters in South America, which incurred the shift of frequent information exchanges from between US and Western Europe to within South America and with the US. Therefore, it is suggested that the

consecutive cosine similarity is able to detect drastic changes of the information flow patterns regarding the shift of active participating clusters and their interactions.

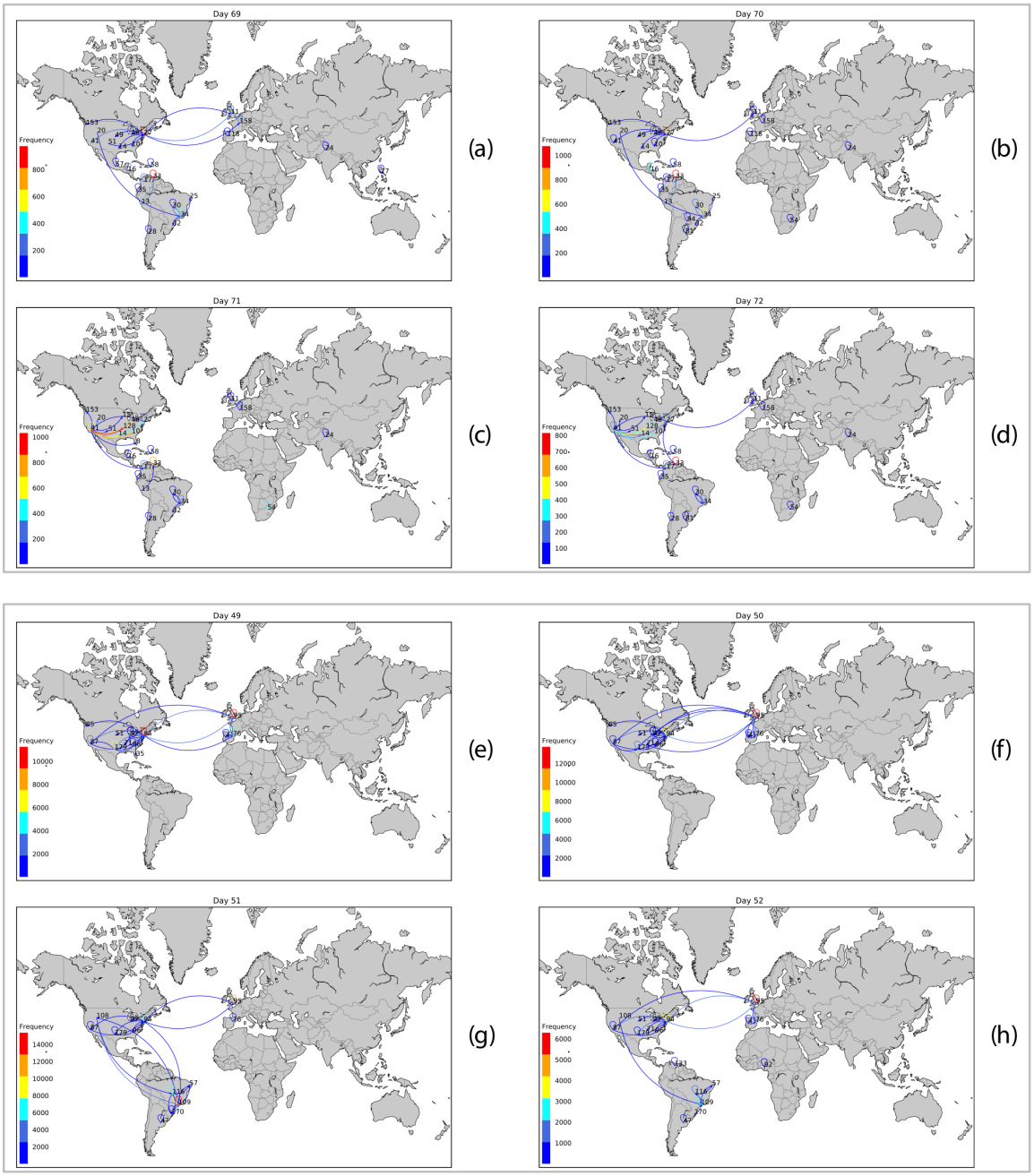

Figure 3.18. Information flow maps of the 69th~72nd days in Zika ((a)–(d)), and 49th~52nd days in Ebola ((e)–(h)).

Next from the dendrograms of pairwise cosine similarity in Figure 3.6 (b) and 3.11 (b), I extracted the grouping results of the 71st day in Zika and 51st day in Ebola. The 71st day in Zika is grouped with a few other days, all in the later stage of the event. It indicates that the 71st day, corresponding to Saturday, February 20th 2016, is a transitioning point of information flow pattern. However in Ebola case, the 51st day corresponding to Friday, October 10th 2014 is the only member of a group, which implies that the 51st day is unique, different from any other days regarding information flow pattern. Therefore, in real-time abnormal pattern detection, the adjacent similarity to its previous day contributes to the discovery of the transition to a new pattern; and the pairwise similarity measure to all its previous days helps unearthing the advent of a unique pattern.

From the dendrograms of pairwise cosine similarity in Figure 3.6 (b) and 3.11 (b), I found that in both events, continuous days are not always grouped together, indicating the constant temporal change of information flow patterns; and discrete days might share similar patterns. Also, I compared the hierarchical clustering results of pairwise cosine similarity shown in Figure 3.6 (c) and 3.11 (c) with the development of daily network properties in Figure 3.14, considering the general trends in segmented phases. A major finding is the mismatch of the grouped days and the days with similar network properties. Therefore, it is suggested that similar network property does not necessarily mean similar information flow pattern, especially when geolocation is considered.

3.5.4.2 Temporal Evolvement of Flow Volume of Top Source and Sink

Daily flow frequencies of the chosen source and sink clusters in Section 3.5.3.2 are summarized and displayed in Figure 3.19. Regarding daily cluster count, the evolvement

of the chosen clusters follows the trend of the overall pattern shown in Figure 3.3 and 3.8. For each chosen cluster, the volume of its inflow and outflow show consistency with time; yet outflow generally surpasses inflow in a major source cluster, and inflow usually exceeds outflow in a major sink cluster. When there are not many participants in the discussion, the difference between inflow and outflow frequencies is relatively small.

I also found that the major source clusters in Figure 3.19 (a) and (e) are normally formed earlier than the major sink clusters in (c) and (g); and the temporal evolvement of the major sink clusters keeps strict consensus with daily point count (Figure 3.3 and 3.8). This is possibly because the formation of sink clusters heavily relies on the number of participants in the discussion, while the construction of source clusters depends on not only the number of users involved in the discussion, but also the developing process of the event. Source clusters are easily shaped surrounding the event location in early stage and some of them have become major source over time, supported by Figure 3.15 and 3.16. At earlier time the information sent from these source clusters are distributed all over the world, which would hinder the build of major sink clusters. At later time with the accumulation of participants, major sink clusters then emerge. In the period of rapid expansion, flow frequencies in both source and sink clusters rise concurrently, and accumulately sink clusters grow more than source clusters.

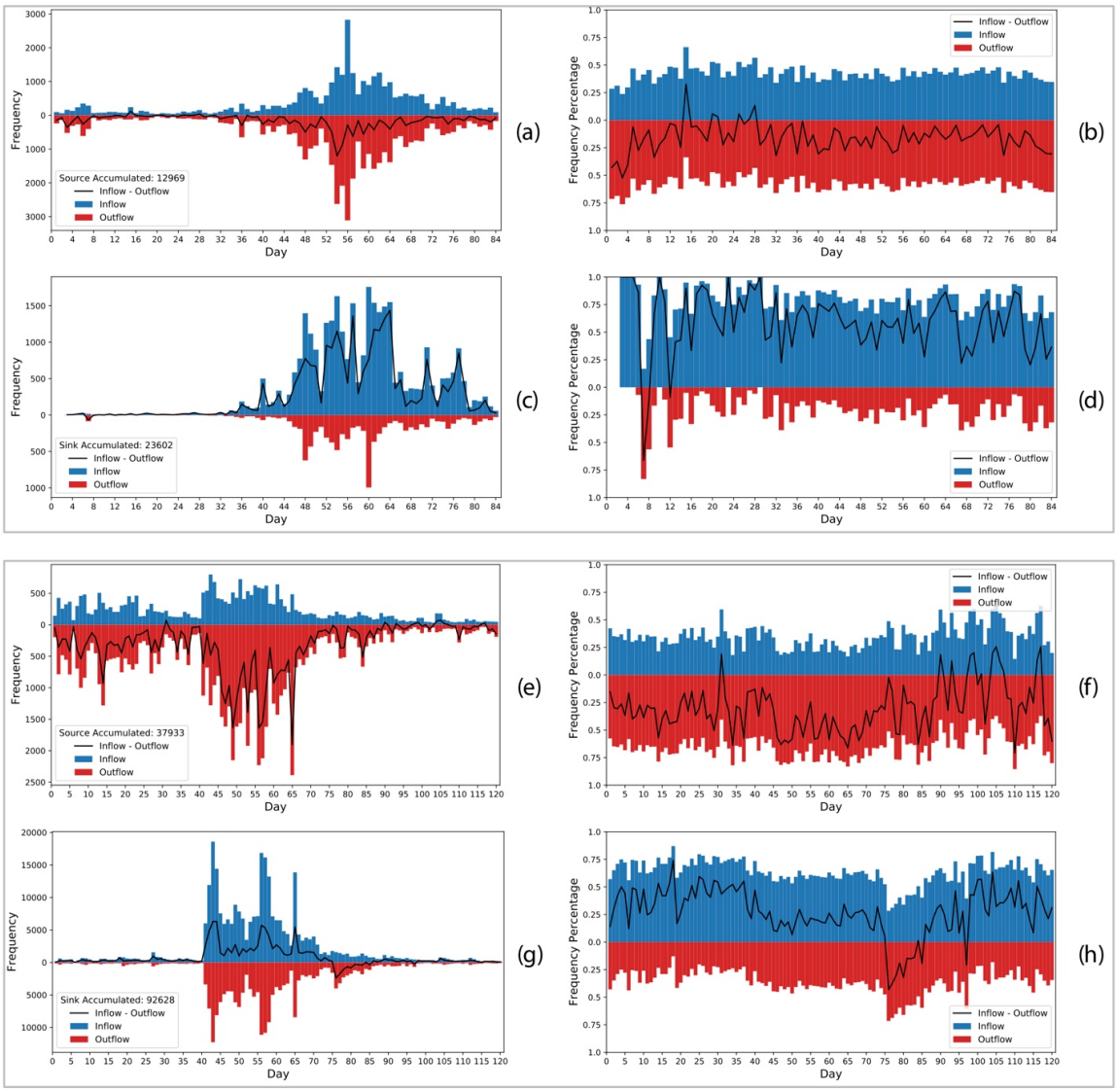

Figure 3.19. Frequency and percentage of frequency of the major source cluster in Brazil in Zika case ((a)–(b)) and West Africa in Ebola case ((e)–(f)), where 'Source Accumulated' means the accumulated difference between outflow and inflow frequencies; and frequency and percentage of frequency of one major sink cluster in the US in Zika case ((c)–(d)) and in Ebola case ((g)–(h)), where 'Sink Accumulated' means the accumulated difference between inflow and outflow frequencies.

3.5.4.3 Temporal Emergence and Disappearance of Source and Sink

As time goes by, new clusters arise and some earlier emerged ones disappear at later time. Figure 3.20-A and 3.20-B illustrate the dynamics of clusters' emergence and disappearance with time in the two event discussions, respectively. Figure 3.20 (a)–(d) are rather self-explanatory, while in (e)–(f), it should be noted that in the label of Y-axis, "changed" refers to any type of change, including new clusters emerging and old clusters disappearing, compared with the previous day. Both figures show very similar patterns and suggest coherent findings. These findings include (1) generally, there are more source clusters that emerge and disappear, comparing to sink clusters on the same day (Figure 3.20-A (a)–(b) and Figure 3.20-B (a)–(b)); (2) the emergence and disappearance of both types of clusters are comparable overall ((c)–(d)); (3) generally source clusters change more significantly than sink clusters especially in later stage ((e)–(f)); and (4) it is more volatile in early days as for clusters change ((a)–(f)).

Unlike the trend of daily point count that shows clear peaks in Figure 3.3 and 3.8, which indicates the participants' join and leaving with the evolvement of events' popularity, the emergence and disappearance of clusters fluctuate with time yet without obvious peaks. The first finding above implies the higher volatility of source clusters; while sink clusters that tend to receive and retweet information are normally more stable.

Besides the second finding above, I also observed greater fluctuation of source clusters in later stage, when the change of sink cluster keeps steady. It suggests that with the fade of interest in the event, participants that have left the discussion would lead to the failure of some source clusters. On the other hand, sink clusters that receive information

still have a buffer period, which keeps the change less radical. This is also supported by the third finding above.

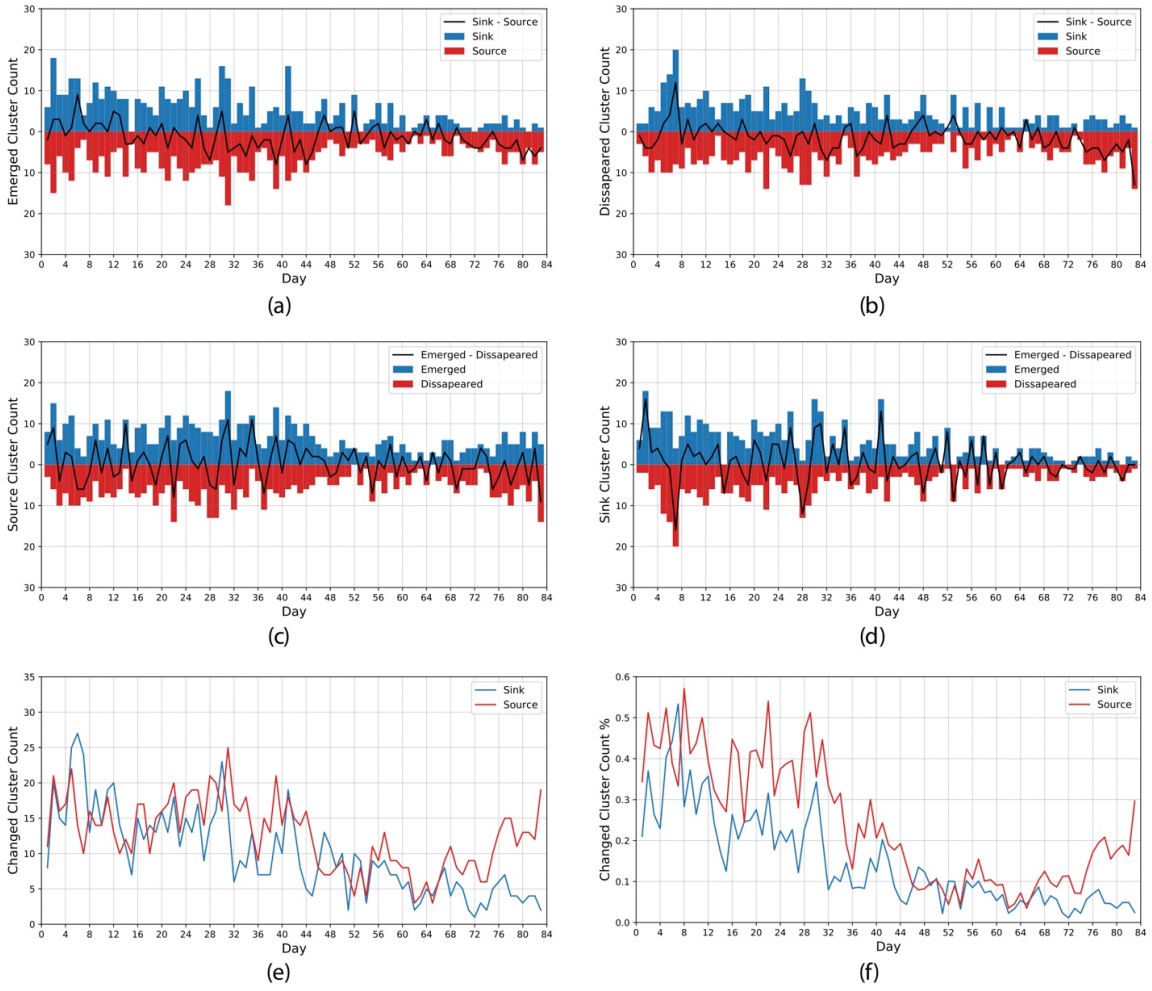

Figure 3.20-A. Daily change of source and sink clusters in Zika case.

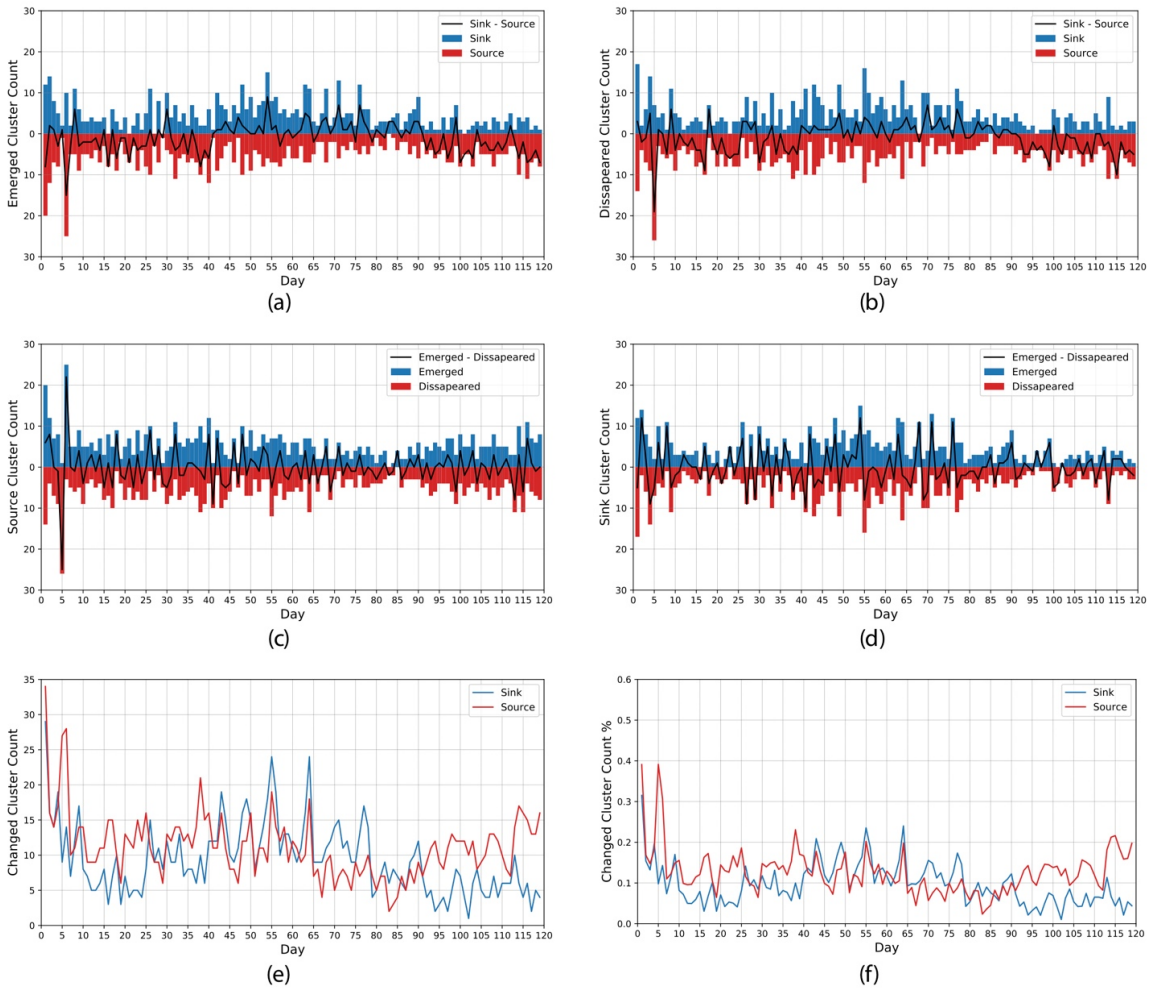

Figure 3.20-B. Daily change of source and sink clusters in Ebola case.

3.6 Conclusions

In this chapter, spatiotemporal analysis was performed on the information diffusion of two epidemics mentioned in Twitter. Stream clustering, information flow analysis, similarity measure, and network analysis were employed. From the obtained results and subsequent discussion, the usefulness of the designed methodology is assured for gaining insights on information diffusion in event discussions over Twitter.

First, the stream clustering method successfully detected clusters that are close in time and space, which laid solid foundation for the following analyses. Based on the adjustment of the spatial and temporal parameters in stream clustering, scale effect on the clustering results is reflected. I argue that the "right scale" to represent the event is largely empirical, depending on the feasibility and our practical demand, such as desired cluster size, stable change of cluster number, minimum abnormality in the similarity measure, and frequency of our inquiry.

Second, network properties, information flow matrices along with their similarity measure, and information flow maps were able to capture the information flow patterns from the structural, quantitative, geographical, and temporal perspectives. Third, from these perspectives of analyses, major social, geographical, and temporal characteristics relevant to an information diffusion process were uncovered. They are: Twitter users' distribution and activeness—closely related to the formation of information sink, which indicates the actual action (i.e. retweeting) for the propagation; geographical scale—directly affects the clusters' size, shape, and containing data points, and further influences the information flow among these clusters; geographical location of the event—shows its influential power especially in the early stage of the discussion, and in the long run it is usually a major information source, providing information to the other places of the world; and time of event progress—in different time periods the flow patterns differ. Future research include testing different temporal resolutions (e.g., weekly) on the same dataset, and applying the designed methodology to different types of events to test its flexibility and robustness.

4 CONCLUSIONS

The emerging field of social media is bursting with new findings that suggest novel social experience. One facet that has drawn my attention is the re-articulation of geographic information (e.g., space, place, and distance) in communication (Graham, 1998). Therefore, this dissertation centers around the communication mechanism in online social networks. It explores the information diffusion process in Twitter's retweet network in the nested cyber and physical environments, in order to help understand the communication pattern of event-associated information in social media platforms.

The increasing availability of Big Data has provided unprecedented opportunities to research on human dynamics and social phenomena (Shaw et al., 2016). However, the use of social media big data in this dissertation has posed challenges, especially in data management and processing. Therefore, facing these challenges, GenDenStream, a stream clustering method was firstly developed. Then based on the clustering results, spatiotemporal analysis was performed on the information diffusion process of public emergencies of international concern widely discussed in Twitter.

Twitter messages were collected for the whole world using Zika and Ebola related keywords suggested by professionals in Public Health. The information diffusion networks through retweeting were examined at the aggregated cluster level. The obtained results and discussion assured the usefulness of the designed methodology for gaining insights on information diffusion in online social networks, and important findings and achievements were gained using the framework:

- The developed GeoDenStream clustering method successfully detected communities that are close in time and space, which has laid solid foundation for the following spatiotemporal analyses. In general, it is particularly suitable for analyzing geotagged social media data streams due to three unique characteristics: its ability to track and maintain information about the identity and composition of clusters over time and space, its ability to handle spatially overlapping data points, and its improved ability to handle noise.
- Analytical methods including network properties, information flow matrices along with their similarity measure, and information flow maps were able to capture the information flow patterns from the structural, quantitative, geographical, and temporal perspectives. Major social, geographical, and temporal characteristics relevant to an information diffusion process were uncovered, including participants' distribution and activeness, geographic scale, geographic location, geography-driven homophily, and time of events progress. This supports the statement that geography matters in the information diffusion process in online social networks.

4.1 Scientific Contributions

By examining the capability of the designed framework and analytical methods, this dissertation research is novel and valuable. Its scientific contributions lie in the advancement of GeoDenStream, the analytical framework for studying information diffusion in online social networks, as well as the discovered information diffusion patterns

along with their potential drives. All guide further research regarding data mining in real-time data streams.

Though the demands for mining key values in Big Data, such as finding and using the hybrid mix of spatial and social contents in social media, have been of great interest and extensively discussed among scientists in various research areas (Croitoru et al., 2017), the common form of its representation, near-continuous data streams (Valle et al., 2009), has brought about tremendous barrier in its handling and analyzing process. Among all efficient operations of this type of data such as basic data analytics, clustering has emerged as one of the most commonly used operations (Krempl et al., 2014; Xu and Tian, 2015). However, the foci of existing clustering methods, such as detecting whether one or more clusters exist and preserving only summary descriptors (e.g., center and radius) of the clusters, are not suitable for our requirement of traceable individual point information and cluster-point relationship in the social media data streams. In view of this requirement, which commonly exists in analyzing geotagged social media data streams, I argue that the proposed GeoDenStream has filled this gap.

In addition to the geotagged social media data streams that are used in this dissertation research, GeoDenStream can be conveniently applied to various application domains as long as the data streams contain geographic information. This is primarily due to the three unique characteristics of GeoDenStream: its ability to track and maintain information about the identity and composition of clusters over time and space, its ability to handle spatially overlapping data points, and its improved ability to handle noise. Potential application domains include but not limited to health, transportation, finance,

energy, climate and weather, and environmental monitoring. Moreover, its capability of handling Big Data is endorsed by the enhanced functionalities and obtained clustering results. The four V's of Big Data are thoroughly responded by (1) stream clustering and memory usage optimization—large volume and high velocity, (2) management of Twitter metadata such as standardizing coordinates information—high variety, and (3) noise cleaning—high veracity.

Online social networks allow internet users to produce, consume, and propagate information at very large scale, and thus have been proved very powerful in information diffusion and influential on society (Guille et al., 2013). In light of this, exploring information diffusion in online social networks is important for understanding the social dynamics and for facilitating higher level reasoning and decision making. This is exactly where the framework of spatiotemporal analysis of information diffusion in this research contributes to.

Analytical methods employed in this research, such as the similarity measure, flow mapping, and network analysis, were effective in revealing the spatiotemporal patterns of information diffusion. For instance, cosine similarity enabled the comparison of flow patterns (i.e. flow volume, and source and sink) at different time steps, and further supported the monitoring of flow patterns change over time and the detection of drastic changes. Flow maps visually illustrated the flow pattern in the geographical layout. Network analysis uncovered the dynamic change of the retweet network properties, which to some extent reflected the composition of participants and their interacting pattern in the Twitter discussions.

In addition to Twitter, data from other social media platforms such as Facebook and Youtube can also utilize this framework for the purpose of exploring online communication that is grounded in physical space. Furthermore, as for research purposes, this framework is particularly useful for detecting drastic change in an event or the occurrence of an abnormal event in general discussion, for identifying major sender and receivers in an information diffusion process for targeted information push, and for spotting moments and periods for effective information propagation.

Potential drives I identified especially the geographical characteristics inspire my curiosity of the external motivation of information diffusion. This is particularly meaningful because in order to manage and make use of information diffusion, it is important to understand what drives this process (Hoang and Mothe, 2018; Suh et al., 2010). Therefore, for higher level reasoning and informed decision making, the important role of the extracted social, geographical, and temporal characteristics in post-internet communication should be reinforced (Kamath et al., 2012). The attributes I detected, such as the geography-related ones—distribution of Twitter users, geolocation of the event, and geography-driven homophily—can serve as the starting point and guidance for further investigation.

As an interdisciplinary research connecting spatiotemporal data mining to social process in the context of online social networks, this dissertation will contribute to varied study fields such as communication, public health, and marketing. The targeted audiences identified in section 1.5 will find this work useful. In addition to the academic field, companies will also benefit from this work. For example, spotting moments and periods

for effective information propagation is applicable to social media marketing; and understanding the related geographical variables to information diffusion gives a spatial view to news and social media practitioners. Further, in the situations when information diffusion is unwanted, for example misinformation, a grasp of this process from all-sided dimensions will help the management and authority parties with effective measures.

4.2 Limitations and Future Work

Despite the novelty and scientific value this dissertation offers, it is still limited in the sole social media data type (i.e., Twitter), the sole research domain it has examined (i.e., public health), and the uncertain relationship between the detected potential drives and an information diffusion process: do they actually drive the diffusion of information, or are they just correlated? In response to these limitations, future research includes expanding the framework beyond Twitter to other types of social media services such as Facebook, applying the designed framework to different application domains, and further examining the role of these potential drives in information diffusion.

More specifically, I propose two examples to illustrate potential application areas. One pertains to the recent abrupt measles outbreaks in the United States. The Center for Disease Control and Prevention (CDC) of US have confirmed 387 individual cases of measles in the first quarter of 2019, already surpassed cases reported of every year except 2014 since measles was eliminated in 2000 (CDC, 2019). A grasp of the online communication pattern of this disease will help us identify highly infected areas, so as to prepare for better and faster countermeasure. Also, it helps to predict the population and area at high risk, so that whom and where vaccination is mostly needed can be targeted

more accurately. The second example focuses on abnormal events. Learning the diffusion process of an abnormal event helps to prevent its reoccurrence in other places, and to better react to its aftermath such as a consequential riot.

In addition to the conceptual discussion of potential drives of the information diffusion through inferences from obtained results and real-world sourced information, quantitative assessment of the potential drives with regards to their influence on information diffusion needs to be further explored. This will facilitate the transformation of our observations to real-world knowledge for informed decision-making. It can be achieved by building regression models that are able to detect useful predicting variables (e.g., Hoang and Mothe, 2018). Variables indicating potential drives can be derived from the tweet records (e.g., time), retweeting network (e.g., network properties of a tweet), and the geographical environment (e.g., country).

Furthermore, an important element in online communication that is overlooked in this research is the content of the information, since my focus is to disclose "how" rather than "what" information diffuses in online social networks. Though this being said, the author is not oblivious of the ample potential of new-found knowledge when content analysis is incorporated to information diffusion. In fact, peer work focusing on this aspect has been done in recent years. For example, researchers have examined the relationship of emotions and information diffusion in social media (Stieglitz and Dang-Xuan, 2013), studied the identification of influential users and relevant content in information diffusion (Silva et al., 2013), and analyzed the spread of low-credibility content by social bots (Shao et al., 2018). Building on the framework and findings of this dissertation research, and by

taking into account content analysis, new knowledge such as the spatiotemporal characteristics of the text content and the relationship between content and the possibility of being spread, can be achieved. This new dimension will undoubtedly enrich the spatiotemporal portrait of information diffusion in online social networks.

REFERENCES

Adamic, L.A., and Glance, N. (2005). The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. In *Proceedings of the 3rd International Workshop on Link Discovery,* (New York, NY, USA: ACM), 36–43.

Adams, P.C. (2011). A taxonomy for communication geography. *Prog. Hum. Geogr. 35*, 37–57.

Agarwal, A., Singh, R., and Toshniwal, D. (2018). Geospatial sentiment analysis using twitter data for UK-EU referendum. *J. Inf. Optim. Sci. 39*, 303–317.

Agarwal, S., Kumar, S., and Goel, U. (2019). Stock market response to information diffusion through internet sources: A literature review. *Int. J. Inf. Manag. 45*, 118–131.

Aggarwal, C.C., Han, J., Wang, J., and Yu, P.S. (2003). A Framework for Clustering Evolving Data Streams. In *Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29,* (Berlin, Germany: VLDB Endowment), 81–92.

Althouse, B.M., Scarpino, S.V., Meyers, L.A., Ayers, J.W., Bargsten, M., Baumbach, J., Brownstein, J.S., Castro, L., Clapham, H., Cummings, D.A., et al. (2015). Enhancing disease surveillance with novel data streams: challenges and opportunities. *EPJ Data Sci. 4*, 17.

Amini, A., Wah, T.Y., and Saboohi, H. (2014). On Density-Based Data Streams Clustering Algorithms: A Survey. J. *Comput. Sci. Technol. 29*, 116–141.

Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J., and Tiwari, M. (2015). Global Diffusion via Cascading Invitations: Structure, Growth, and Homophily. In *Proceedings of the 24th International Conference on World Wide Web, (Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee)*, 66–76.

Aral, S., Muchnik, L., and Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci. 106*, 21544–21549.

Atefeh, F., and Khreich, W. (2015). A Survey of Techniques for Event Detection in Twitter. *Comput. Intell. 31*, 132–164.

Atluri, G., Karpatne, A., & Kumar, V. (2018). Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys (CSUR)*, *51*(4), 83.

Bakshy, E., Hofman, J.M., Mason, W.A., and Watts, D.J. (2011). Everyone's an Influencer: Quantifying Influence on Twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining,* (New York, NY, USA: ACM), 65–74.

Bello-Orgaz, G., Jung, J.J., and Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Inf. Fusion 28*, 45–59.

Bifet, A., Holmes, G., Kirkby, R., and Pfahringer, B. (2010). MOA: Massive Online Analysis. J. Mach. Learn. Res. *11*, 1601–1604.

Boccaletti, S., Bianconi, G., Criado, R., del Genio, C.I., Gómez-Gardeñes, J., Romance, M., Sendiña-Nadal, I., Wang, Z., and Zanin, M. (2014). The structure and dynamics of multilayer networks. *Phys. Rep. 544*, 1–122.

Borgatti, S.P. (2005). Centrality and network flow. *Soc. Netw. 27*, 55–71.

Boyd, D., Golder, S., and Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *2010 43rd Hawaii International Conference on System Sciences,* 1–10.

Budak, C., Agrawal, D., and El Abbadi, A. (2011). Limiting the Spread of Misinformation in Social Networks. In *Proceedings of the 20th International Conference on World Wide Web,* (New York, NY, USA: ACM), 665–674.

Cao, F., Estert, M., Qian, W., and Zhou, A. (2006). Density-Based Clustering over an Evolving Data Stream with Noise. In *Proceedings of the 2006 SIAM International Conference on Data Mining, (Society for Industrial and Applied Mathematics)*, 328–339.

CDC (2019). Measles Cases in 2019 (Center for Disease Control and Prevention).

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K.P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. In *Fourth International AAAI Conference on Weblogs and Social Media.*

Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D.S., and Ertl, T. (2012). Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST),* 143–152.

Chen, M., Peng, Y., Li, A., Li, Z., Deng, Y., Liu, W., Liao, B., and Dai, C. (2018a). A novel information diffusion method based on network consistency for identifying disease related microRNAs. RSC Adv. *8*(64), 36675–36690.

Chen, Y., Huang, Z., Pei, T., and Liu, Y. (2018b). HiSpatialCluster: A novel high-performance software tool for clustering massive spatial points. *Trans. GIS, 22*, 1275–1298.

CNN (2009). Facebook flashmob shuts down station. *CNN.com*. Retrieved from: http://www.cnn.com/2009/WORLD/europe/02/09/uk.station.flashmob/index.html

Cox, J.B. (2016). News orgs post more often on Twitter than on Facebook. *Newsp. Res. J. 37*, 220–234.

Croitoru, A., Crooks, A., Radzikowski, J., and Stefanidis, A. (2013). Geosocial gauge: a system prototype for knowledge discovery from social media. *Int. J. Geogr. Inf. Sci. 27*, 2483–2508.

Croitoru, A., Crooks, A.T., Radzikowski, J., Stefanidis, A., Vatsavai, R.R., and Wayant, N. (2014). Geoinformatics and Social Media New Big Data Challenge.

Croitoru, A., Wayant, N., Crooks, A., Radzikowski, J., and Stefanidis, A. (2015). Linking cyber and physical spaces through community detection and clustering in social media feeds. *Comput. Environ. Urban Syst. 53*, 47–64.

Croitoru, A., Crooks, A., Radzikowski, J., and Stefanidis, A. (2017). Geovisualization of Social Media. Int. Encycl. Geogr*: People, the Earth, Environment and Technology*, 1-17.

Crooks, A.T., Hudson-Smith, A., Croitoru, A., and Stefanidis, A. (2014). The Evolving GeoWeb. In *Geocomputation*, (Boca Raton: FL: CRC Press), 67–94.

Crooks, A.T., Croitoru, A., Jenkins, A., Mahabir, R., Agouris, P., and Stefanidis, A. (2016). User-Generated Big Data and Urban Morphology. *Built Environment*, *42*(3), 396-414.

De Choudhury, M., Lin, Y.-R., Sundaram, H., Candan, K.S., Xie, L., and Kelliher, A. (2010). How Does the Data Sampling Strategy Impact the Discovery of Information Diffusion in Social Media? *ICWSM* 34–41.

Ek, R. (2006). Media Studies, Geographical Imaginations and Relational Space. *Geogr. Commun. Spat. Turn Media Stud.* 45–66.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining,* (Portland, Oregon: AAAI Press), 226–231.

Estivill-Castro, V. (2002). Why So Many Clustering Algorithms: *A Position Paper. SIGKDD Explor Newsl 4*, 65–75.

Everitt, B.S., and Skrondal, A. (2010). The Cambridge Dictionary of Statistics (New York: Cambridge University Press).

Falconer, W. (1781). Remarks on the Influence of Climate, Situation, Nature of Country, Population, Nature of Food and Way of Life on the Disposition and Temper, Manners and Behaviour, Intellects, Laws and Customs, Form of Goverment End Religion of Mankind (C. Dilly).

Ferrara, E., Varol, O., Menczer, F., and Flammini, A. (2013). Traveling Trends: Social Butterflies or Frequent Fliers? In *Proceedings of the First ACM Conference on Online Social Networks,* (New York, NY, USA: ACM), 213–222.

Flatow, D., Naaman, M., Xie, K.E., Volkovich, Y., and Kanza, Y. (2015). On the Accuracy of Hyper-local Geotagging of Social Media Content. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining,* (New York, NY, USA: ACM), 127–136.

Forestiero, A., Pizzuti, C., and Spezzano, G. (2009). FlockStream: A Bio-Inspired Algorithm for Clustering Evolving Data Streams. In *2009 21st IEEE International Conference on Tools with Artificial Intelligence(ICTAI)*, 1–8.

Freeman, E., Woodruff, S.D., Worley, S.J., Lubker, S.J., Kent, E.C., Angel, W.E., Berry, D.I., Brohan, P., Eastman, R., Gates, L., et al. (2017). ICOADS Release 3.0: a major update to the historical marine climate record. *Int. J. Climatol. 37*, 2211–2232.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., et al. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Sci. Data 2*, 150066.

Gaber, M.M., Zaslavsky, A., and Krishnaswamy, S. (2005). Mining Data Streams: A Review. *SIGMOD Rec 34*, 18–26.

Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z., and Kellerer, W. (2010). Outtweeting the Twitterers - Predicting Information Cascades in Microblogs. In *Proceedings of the 3rd Conference on Online Social Networks,* (Berkeley, CA, USA: USENIX Association), 3–3.

Gartner (2018). Big Data. Gart. IT Gloss. Retrieved from: https://www.gartner.com/it-glossary/big-data/

Ghesmoune, M., Lebbah, M., and Azzag, H. (2016). State-of-the-art on clustering data streams. *Big Data Anal. 1*, 13.

Gomez-Rodriguez, M., Leskovec, J., and Krause, A. (2012). Inferring Networks of Diffusion and Influence. *ACM Trans Knowl Discov Data 5*, 21:1–21:37.

Goodchild, M.F. (2007). Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *Int. J. Spat. Data Infrastruct. Res. 2*, 24–32.

Gottfried, J., and Shearer, E. (2016). News Use Across Social Media Platforms 2016.

Graham, S. (1998). The end of geography or the explosion of place? Conceptualizing space, place and information technology. *Prog. Hum. Geogr. 22*, 165–185.

Graham, S., and Marvin, S. (1995). Telecommunications and the City: Electronic Spaces, Urban Places (London ; New York: Routledge).

Guille, A., Hacid, H., Favre, C., and Zighed, D.A. (2013). Information Diffusion in Online Social Networks: A Survey. *SIGMOD Rec 42*, 17–28.

Hahsler, M., Bolaños, M., and Forrest, J. (2015). streamMOA: Interface for MOA Stream Clustering Algorithms. R Package Version 1.

Hahsler, M., Bolaños, M., and Forrest, J. (2017). Introduction to stream: An Extensible Framework for Data Stream Clustering Research with R. *J. Stat. Softw. 76*, 1–50.

Hale, S.A. (2014). Global Connectivity and Multilinguals in the Twitter Network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* (New York, NY, USA: ACM), 833–842.

Han, J., Pei, J., and Kamber, M. (2011). Data Mining: Concepts and Techniques (Elsevier).

Hannigan, J., Hernandez, G., Medina, R.M., Roos, P., and Shakarian, P. (2013). Mining for Spatially-Near Communities in Geo-Located Social Networks. In *2013 AAAI Fall Symposium Series*.

Harvey, D. (1984). On the History and Present Condition of Geography: An Historical Materialist Manifesto. *Prof. Geogr. 36*, 1–11.

Hassani, M., Spaus, P., Gaber, M.M., and Seidl, T. (2012). Density-Based Projected Clustering of Data Streams. In *Scalable Uncertainty Management, E. Hüllermeier, S. Link, T. Fober, and B. Seeger, eds.* (Springer Berlin Heidelberg), 311–324.

Hoang, T.B.N., and Mothe, J. (2018). Predicting information diffusion on Twitter – Analysis of predictive features. *J. Comput. Sci. 28*, 257–264.

Hoffman, S.J., and Silverberg, S.L. (2018). Delays in Global Disease Outbreak Responses: Lessons from H1N1, Ebola, and Zika. *Am. J. Public Health 108*, 329–333.

Isaksson, C., Dunham, M.H., and Hahsler, M. (2012). SOStream: Self Organizing Density-Based Clustering over Data Stream. In Machine Learning and Data Mining in Pattern Recognition, *P. Perner, ed.* (Springer Berlin Heidelberg), 264–278.

Jacobsen, K.H., Aguirre, A.A., Bailey, C.L., Baranova, A.V., Crooks, A.T., Croitoru, A., Delamater, P.L., Gupta, J., Kehn-Hall, K., Narayanan, A., et al. (2016). Lessons from the Ebola Outbreak: Action Items for Emerging Infectious Disease Preparedness and Response. *EcoHealth 13*, 200–212.

Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis,* (New York, NY, USA: ACM), 56–65.

Jenkins, A., Croitoru, A., Crooks, A.T., and Stefanidis, A. (2016). Crowdsourcing a Collective Sense of Place. *PloS One 11*, e0152932.

Kaisler, S., Armour, F., Espinosa, J.A., and Money, W. (2013). Big Data: Issues and Challenges Moving Forward. In *2013 46th Hawaii International Conference on System Sciences,* 995–1004.

Kamath, K.Y., Caverlee, J., Cheng, Z., and Sui, D.Z. (2012). Spatial Influence vs. Community Influence: Modeling the Global Spread of Social Media. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, (New York, NY, USA: ACM), 962–971.

Kaplan, A.M., and Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Bus. Horiz. 53*, 59–68.

Kietzmann, J.H., Hermkens, K., McCarthy, I.P., and Silvestre, B.S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Bus. Horiz. 54*, 241–251.

Kim, H., Beznosov, K., and Yoneki, E. (2014). Finding Influential Neighbors to Maximize Information Diffusion in Twitter. In *Proceedings of the 23rd International Conference on World Wide Web,* (New York, NY, USA: ACM), 701–706.

Kim, J., Bae, J., and Hastak, M. (2018). Emergency information diffusion on online social media during storm Cindy in U.S. *Int. J. Inf. Manag. 40*, 153–165.

Kindhauser, M.K., Allen, T., Frank, V., Santhana, R.S., and Dye, C. (2016). Zika: the origin and spread of a mosquito-borne virus. *Bulletin of the World Health Organization*, *94*(9), 675.

Kranen, P., Kremer, H., Jansen, T., Seidl, T., Bifet, A., Holmes, G., Pfahringer, B., and Read, J. (2012). Stream Data Mining Using the MOA Framework. In *Database Systems for Advanced Applications,* S. Lee, Z. Peng, X. Zhou, Y.-S. Moon, R. Unland, and J. Yoo, eds. (Springer Berlin Heidelberg), 309–313.

Krempl, G., Žliobaite, I., Brzeziński, D., Hüllermeier, E., Last, M., Lemaire, V., Noack, T., Shaker, A., Sievi, S., Spiliopoulou, M., et al. (2014). Open Challenges for Data Stream Mining Research. *SIGKDD Explor Newsl 16*, 1–10.

Kromwijk, K., Balkesen, Ç., Boder, G., Dindar, N., Keusch, F., Sengül, A., and Tatbul, N. (2010). Connecting the Real World with the Virtual World: The SmartRFLib RFID-Supported Library System on Second Life. *Handb. Res. Web 20 30 X0 Technol. Bus. Soc. Appl.* 720–732.

Kulshrestha, J., Kooti, F., Nikravesh, A., & Gummadi, K. P. (2012, May). Geographic dissection of the twitter network. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Kwon, K.H., Wang, H., Raymond, R., and Xu, W.W. (2015). A Spatiotemporal Model of Twitter Information Diffusion: An Example of Egyptian Revolution 2011. In *Proceedings of the 2015 International Conference on Social Media & Society*, (New York, NY, USA: ACM), 4:1–4:7.

La Fond, T., and Neville, J. (2010). Randomization Tests for Distinguishing Social Influence and Homophily Effects. In *Proceedings of the 19th International Conference on World Wide Web*, (New York, NY, USA: ACM), 601–610.

Lerman, K. (2007). Social Information Processing in News Aggregation. *IEEE Internet Comput*. *11*, 16–28.

Li, P., Lu, H., Kanhabua, N., Zhao, S., and Pan, G. (2018a). Location Inference for Non-geotagged Tweets in User Timelines. *IEEE Trans. Knowl. Data Eng.* 1–1.

Li, X., Cheng, X., Su, S., and Sun, C. (2018b). Community-based seeds selection algorithm for location aware influence maximization. *Neurocomputing 275*, 1601–1613.

Lin, J., and Lin, H. (2009). A density-based clustering over evolving heterogeneous data stream. In *2009 ISECS International Colloquium on Computing, Communication, Control, and Management,* 275–277.

Liu, L., Huang, H., Guo, Y., and Chen, F. (2009). rDenStream, A Clustering Algorithm over an Evolving Data Stream. In *2009 International Conference on Information Engineering and Computer Science*, 1–4.

Liu, W., Zheng, Y., Chawla, S., Yuan, J., and Xing, X. (2011). Discovering Spatio-temporal Causal Interactions in Traffic Data Streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* (New York, NY, USA: ACM), 1010–1018.

Liu, X., Wu, X., Wang, H., Zhang, R., Bailey, J., and Ramamohanarao, K. (2010). Mining distribution change in stock order streams. In *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, 105–108.

Lotan, G. (2011). Mapping Information Flows on Twitter. *Future Soc. Web 5*.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., and Boyd, D. (2011). The Arab Spring| The Revolutions Were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions. *Int. J. Commun. 5*, 31.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comput. 17*, 395–416.

Malik, M.M., Lamba, H., Nakos, C., and Pfeffer, J. (2015). Population Bias in Geotagged Tweets. In *Ninth International AAAI Conference on Web and Social Media,* 18–27.

Matsa, K., and Shearer, E. (2018). News Use Across Social Media Platforms 2018 | Pew Research Center. Retrieved from http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/

McPherson, M., Smith-Lovin, L., and Cook, J.M. (2001). Birds of a Feather: Homophily in Social Networks. *Annu. Rev. Sociol. 27*, 415–444.

Mishori, R., Singh, L.O., Levy, B., and Newport, C. (2014). Mapping Physician Twitter Networks: Describing How They Work as a First Step in Understanding Connectivity, Information Flow, and Message Diffusion. *J. Med. Internet Res. 16*.

Mislove, A., Viswanath, B., Gummadi, K.P., and Druschel, P. (2010). You Are Who You Know: Inferring User Profiles in Online Social Networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining,* (New York, NY, USA: ACM), 251–260.

Mitchell, W.J. (1996). City of Bits: Space, Place, and the Infobahn (Cambridge, Mass.: The MIT Press).

Mitchell, A., Barthel, M., Shearer, E., and Gottfried, J. (2015). The Evolving Role of News on Twitter and Facebook. *Pew Research Center, 14*.

Mok, D., Wellman, B., and Carrasco, J.-A. (2010). Does Distance Matter in the Age of the Internet? *Urban Stud. 47*, 2747–2783.

Molaei, S., Zare, H., and Veisi, H. (2019). Deep Learning Approach on Information Diffusion in Heterogeneous Networks. *arXiv preprint arXiv:1902.08810*.

Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J., and Damas, L. (2016). Time-evolving O-D matrix estimation using high-speed GPS data streams. *Expert Syst. Appl. 44*, 275–288.

Mouratidis, K., Papadias, D., and Hadjieleftheriou, M. (2005). Conceptual Partitioning: An Efficient Method for Continuous Nearest Neighbor Monitoring. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data,* (New York, NY, USA: ACM), 634–645.

Naaman, M., Boase, J., and Lai, C.-H. (2010). Is It Really About Me?: Message Content in Social Awareness Streams. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work,* (New York, NY, USA: ACM), 189–192.

Ntoutsi, I., Zimek, A., Palpanas, T., Kröger, P., and Kriegel, H. (2012). Density-based Projected Clustering over High Dimensional Data Streams. In *Proceedings of the 2012 SIAM International Conference on Data Mining,* (Society for Industrial and Applied Mathematics), 987–998.

O'Callaghan, L., Mishra, N., Meyerson, A., Guha, S., and Motwani, R. (2002). Streaming-data algorithms for high-quality clustering. In *Proceedings 18th International Conference on Data Engineering,* 685–694.

O'Reilly, T. (2007). What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software (University Library of Munich, Germany).

Palpanas, T., and Paraskevopoulos, P. (2015). Fine-grained geolocalisation of non-geotagged tweets. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM),* 105–112.

PeerReach (2013). 4 ways how Twitter can keep growing. *PeerReach Blog*.

Powers, D.M. (2011). Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation.

Pruthi, P., Yadav, A., Abbasi, F., and Toshniwal, D. (2015). How Has Twitter Changed the Event Discussion Scenario? A Spatio-temporal Diffusion Analysis. In *2015 IEEE International Congress on Big Data,* (New York City, NY, USA: IEEE), 733–736.

Puri, A., Arora, P., and Sardana, N. (2018). Analysis and Visualisation of Geo-Referenced Tweets for Real-Time Information Diffusion. *Procedia Comput. Sci. 132*, 1138–1146.

Ramezani, M., Khodadadi, A., and Rabiee, H.R. (2018). Community Detection Using Diffusion Information. *ACM Trans Knowl Discov Data 12*, 20:1–20:22.

Relph, E. (1976). Place and Placelessness (London: Pion Ltd).

Ren, J., and Ma, R. (2009). Density-Based Data Streams Clustering over Sliding Windows. In *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery,* 248–252.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math. 20*, 53–65.

Ruiz, C., Menasalvas, E., and Spiliopoulou, M. (2009). C-DenStream: Using Domain Knowledge on a Data Stream. In *Discovery Science,* (Springer, Berlin, Heidelberg), 287–301.

Saito, K., Kimura, M., Ohara, K., and Motoda, H. (2010). Selecting Information Diffusion Models over Social Networks for Behavioral Analysis. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III,* (Berlin, Heidelberg: Springer-Verlag), 180–195.

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., and Tufano, P. (2012). Analytics: The Real-World Use of Big Data, *IBM report*.

Shaban, H. (2019). Twitter reveals its daily active user numbers for the first time. Wash. Post. Retrieved from: https://www.washingtonpost.com/technology/2019/02/07/twitter-reveals-its-daily-active-user-numbers-first-time/

Shao, C., Ciampaglia, G.L., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. *Nat. Commun. 9*, 4787.

Shaw, S.-L., Tsou, M.-H., and Ye, X. (2016). Editorial: human dynamics in the mobile and big data era. *Int. J. Geogr. Inf. Sci. 30*, 1687–1693.

Shearer, E. (2015). 5 key takeaways about Twitter, Facebook and news use. *Pew Research Center*. Retrived from: https://www.pewresearch.org/fact-tank/2015/07/14/5-key-takeaways-about-twitter-facebook-and-news-use/

Silva, A., Guimarães, S., Meira, W., Jr., and Zaki, M. (2013). ProfileRank: Finding Relevant Content and Influential Users Based on Information Diffusion. In

Proceedings of the 7th Workshop on Social Network Mining and Analysis, (New York, NY, USA: ACM), 2:1–2:9.

Simons-Morton, B., and Farhat, T. (2010). Recent Findings on Peer Group Influences on Adolescent Substance Use. *J. Prim. Prev. 31*, 191–208.

Skála, J., and Kolingerová, I. (2011). Dynamic hierarchical triangulation of a clustered data stream. *Comput. Geosci. 37*, 1092–1101.

Soja, E.W. (1985). The Spatiality of Social Life: Towards a Transformative Retheorisation. In *Social Relations and Spatial Structures, D. Gregory, and J. Urry, eds.* (Macmillan Education UK), 90–127.

Stefanidis, A., Crooks, A., and Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal, 78*, 319–338.

Stefanidis, A., Vraga, E., Lamprianidis, G., Radzikowski, J., Delamater, P.L., Jacobsen, K.H., Pfoser, D., Croitoru, A., and Crooks, A. (2017). Zika in Twitter: Temporal Variations of Locations, Actors, and Concepts. *JMIR Public Health Surveill. 3*, 22.

Stieglitz, S., and Dang-Xuan, L. (2013). Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior. *J. Manag. Inf. Syst. 29*, 217–248.

Suh, B., Hong, L., Pirolli, P., and Chi, E.H. (2010). Want to Be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing,* (Washington, DC, USA: IEEE Computer Society), 177–184.

Tan, Z., Wu, D., Gao, T., You, I., and Sharma, V. (2019). AIM: Activation increment minimization strategy for preventing bad information diffusion in OSNs. *Future Gener. Comput. Syst. 94*, 293–301.

Tasoulis, D.K., Ross, G., and Adams, N.M. (2007). Visualising the Cluster Structure of Data Streams. In *Advances in Intelligent Data Analysis VII, M. R. Berthold, J. Shawe-Taylor, and N. Lavrač, eds.* (Springer Berlin Heidelberg), 81–92.

Tobler, W.R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Econ. Geogr. 46*, 234–240.

Tuan, Y.-F. (2001). Space and Place: The Perspective of Experience. *Minneapolis, Minn.: Univ Of Minnesota Press*.

Valle, E.D., Ceri, S., Harmelen, F. v, and Fensel, D. (2009). It's a Streaming World! Reasoning upon Rapidly Changing Information. *IEEE Intell. Syst. 24*, 83–89.

Vikhorev, K., Greenough, R., and Brown, N. (2013). An advanced energy management framework to promote energy awareness. *J. Clean. Prod. 43*, 103–112.

Vinh, N.X., Epps, J., and Bailey, J. (2010). Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *J. Mach. Learn. Res. 11*, 2837–2854.

Wang, F., Jiang, W., Li, X., and Wang, G. (2018). Maximizing positive influence spread in online social networks via fluid dynamics. *Future Gener. Comput. Syst. 86*, 1491–1502.

Wellman, B. (2001). Little Boxes, Glocalization, and Networked Individualism. In Revised Papers from the Second Kyoto Workshop on Digital Cities II, *Computational and Sociological Approaches,* (London, UK, UK: Springer-Verlag), 10–25.

WHO (2016). Zika situation report (WHO). Retrieved from: http://www.who.int/emergencies/zika-virus/situation-report/1-september-2016/en/

WHO (2018). 2014-2016 Ebola Outbreak in West Africa. *CDC*. Retrieved from https://www.cdc.gov/vhf/ebola/history/2014-2016-outbreak/index.html

Woo, J., and Chen, H. (2016). Epidemic model for information diffusion in web forums: experiments in marketing exchange and political dialog. *SpringerPlus 5*, 66.

Xu, D., and Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. *Ann. Data Sci. 2*, 165–193.

Yerasani, S., Appam, D., Sarma, M., and Tiwari, M.K. (2019). Estimation and maximization of user influence in social networks. *Int. J. Inf. Manag. 47*, 44–51.

Yin, L., Kretschmer, H., Hanneman, R.A., and Liu, Z. (2006). Connection and stratification in research collaboration: An analysis of the COLLNET network. *Inf. Process. Manag. 42*, 1599–1613.

Zafarani, R., Abbasi, M.A., and Liu, H. (2014). Social Media Mining: An Introduction (New York, NY: Cambridge University Press).

Zarrella, D. (2009). The Science Of Retweets. *Business*. Retrieved from https://www.slideshare.net/danzarrella/the-science-of-re-tweets.

Zhang, W., Ye, Y., Tan, H., Dai, Q., and Li, T. (2013). Information Diffusion Model Based on Social Network. In *Proceedings of the 2012 International Conference of Modern Computer Science and Applications*, (Springer, Berlin, Heidelberg), 145–150.

Zhang, Z.-K., Liu, C., Zhan, X.-X., Lu, X., Zhang, C.-X., and Zhang, Y.-C. (2016). Dynamics of information diffusion and its applications on complex networks. *Phys. Rep. 651*, 1–34.

Zhao, N., and Li, H. (2019). How can social commerce be boosted? The impact of consumer behaviors on the information dissemination mechanism in a social commerce network. *Electron. Commer. Res*., 1–24.

Zhu, X., Hao, J., Shen, Y., Liu, T., and Liu, M. (2018). Diffusion of False Information During Public Crises: Analysis Based on the Cellular Automaton Method. *Comput. Inform. 37*, 23-48–48.

OutlierDenStream (2018). OutlierDenStream. Retrived from https://github.com/anrputina/OutlierDenStream