Communications

A Topological Data Analysis Approach to Visualizing Ebola Tweets

Herchel Thaddeus Machacon [*1]

The role of social media in health and medical information in general, and during an epidemic in particular, has been reported. Data from social media such as tweets provide vast opportunities and potential benefits for health and medical information communication. However, openly-available social media network visualization tools focus on the online connections between social media users which may not be of utmost importance for health and medical information practitioners. This paper takes the topological data analysis (TDA) approach to render a visual representation of these large, unstructured, and highly complex data. Utilizing a TDA and machine learning platform, distinct features of these Ebola tweets were visually and statistically identified.

Topological locations of relevant keywords (*virus, epidemic, Africa, Sierra Leone, blood, saliva, fever, misinformation,* etc.) contained in these tweets, as well as the topological locations of news and health-related organizations are presented.

Key words: Ebola tweets, Social media, Visualization, Topological data analysis

1. Introduction

The 2014 Ebola epidemic that affected several countries in West Africa was reportedly the largest in history. Over the span of a year, the Ebola epidemic has caused more than 10 times as many cases of Ebola than the combined total of all those reported in previous Ebola outbreaks[1)]. Prolific coverage of the news in the mainstream as well as in social media followed[2)]. From July to August 2014, during the early stages of the 2014 Ebola outbreak, there were 42,236 tweets mentioning Ebola[3)]. In October 2014, there were more than 21 million tweets about Ebola[4)]. In January 2016, BBC reported a new case of Ebola in Sierra Leone[5)].

There is an abundance of literature attesting to the potential of social networking sites such as Twitter not only as tools to disseminate health-related information but also as tools for the prediction and surveillance of diseases[6~12)].

Tweets are highly unstructured text data containing grammatical and spelling errors, non-alphabet characters including emoji or ideograms. Twitter reportedly generates 500 million tweets per day and is projected to reach 1.2 billion

[*1]桐生大学 医療保健学部
〒379-2932　みどり市笠懸町阿左美 606-7
E-mail：machacon-htc@kiryu-u.ac.jp
受付日：2016 年 1 月 25 日
採択日：2016 年 10 月 27 日

[*1]Kiryu University, Faculty of Health Care
606-7 Kasakakemachi Azami, Gunma, 379-2932, Japan

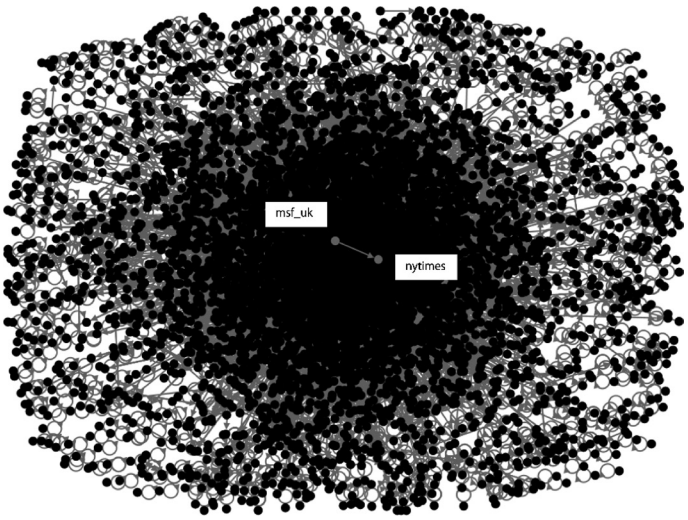

Fig. 1 Ebola Tweet Network Plotted with NodeXL
Nodes represent vertices whose tweets contain the keyword "ebola", mentions or replies-to other vertices. The "*who-mentions-who*" or "*who-replies to-whom*" relationship between *msf_uk* and *nytimes* is illustrated above. Here, *msf_uk* follows the *nytimes*.

tweets per day, 1.36 petabytes per year by 2025[13].

However, information visualization tools are needed to harvest opportunities provided by this massive amount of textual data. Interactive visualization methods can help to improve healthcare[14] and to monitor infectious disease outbreaks[15]. Visualization tools such as tweet maps[16], trend graphs[4,17], tag clouds of keywords[9], as well as social network diagrams[14,18,19] have been developed and used to visualize data mined from Twitter. One of the tools used to visualize social media networks and their underlying connections is NodeXL[14,20,21] (http://nodexl.codeplex.com). NodeXL plots social media network data as vertices and edges. In other words, it builds a network structure plotting each twitter user or organization as a single vertex with edges signifying their connections or relationships with other tweeters. **Figure 1** shows the Ebola tweet network plotted with NodeXL. Nodes represent vertices whose tweets contain the keyword "Ebola", as well as those who mentions or replies-to these vertices. The network was plotted using the Fructerman-Reingold force-directed layout algorithm which is known for its speed and robustness. The relationship between two vertices, *msf-uk* (Médecins Sans Frontiéres https://twitter.com/msf_uk) and the *nytimes* (The New York Times https://twitter.com/nytimes) is illustrated. Here, an arrow which represents the edge points from *msf_uk* to *nytimes*, i.e., *msf_uk* follows *nytimes*. With about 47,000 tweets plotted, the central structure is extremely dense. This accounts only for about 60% of the number of tweet data considered in this study. If the total number of tweets data examined were to be plotted using the same algorithm, one can see nothing but a huge black "hairball" of indiscernible connections. The network shown in **Figure 1** shows only the relationships or connections between vertices.

While social media network visualization tools such as NodeXL focus on the online connections that make up the social media structure, this work focuses not on the "*who-mentions-who*" or "*who-replies to-whom*" relationship, but rather on finding insights or characteristics hidden in these complex and voluminous data. Are there similarities or differences between these data points beyond the categorization of tweet relationship as a tweet, mentions, or replies-to relationship? Can we visually identify and examine distinct features in the Ebola tweet network structure?

Here, the power of Ayasdi software (http://www.ayasdi.com) is leveraged to analyze large and complex data using topological data analysis (TDA) and machine learning. It has been applied to diverse data such as gene expressions from breast tumors, voting data from the United States, and player performance from the NBA[22].

Topology is defined as one of the branches of mathematics which studies properties of shape[23,24]. Topological data analysis (TDA) uses topological methods as a geometric approach to pattern or shape recognition within data. One can thus discover insights from data by recognizing shapes within the data itself. Lum and co-authors[22] argue that TDA's advantage lies in its sensitivity to both large and small scale patterns that other analytical methods such as principal component analysis (PCA), multidimensional scaling (MDS), and cluster analysis often fail to detect. TDA does not depend on prior hypotheses that focus on pairwise relationships as mentioned in the same study. It should also be noted that TDA does not depend on the coordinate system chosen, and as such allows the comparison of data derived from different coordinate systems.

The aim of this paper is to use the TDA approach to render a visual representation of these large, unstructured, and highly complex data, and extract insights from Ebola tweets that may not be apparent using openly-available social media network visualization tools.

2. Methods

Data which comprise tweets containing the keyword Ebola were collected through NodeXL's Search Network which utilizes Twitter's search API. This allows us to collect people or organizations that have included the term Ebola in their tweets and those who mention or replies to these individuals or organizations. Here, the data represent 74,618 tweets covering a period of 4 months from June to September of 2014. It includes two vertices (vertex 1 which is the source of the tweet, vertex 2 which is the destination to which vertex 1 points to), the relationship between these vertices (this could be a tweet, a mention, or a reply-to), the tweet content, domains or URLs contained in the tweets themselves, and hashtags included in the tweet, the tweet date in UTC, the number of users who follow vertex 1 as well as the number of followers vertex 1 has, and the time zones. **Table 1** shows the columns in one data point or row, and the sample content. No data cleaning steps were performed.

Data was uploaded into Ayasdi Core, a platform that combines TDA and machine learning, which has the capability to render a simple visual representation of large and highly complex data and to reveal significant patterns hidden within these data. Ayasdi Core outputs a topological network where nodes do not represent a single data but rather a set of similar data points, and edges that connect nodes with common data points. **Table 2** shows the data dimensions subjected to TDA analysis, i.e. data used

Table 1 Sample Data Mined from Twitter

Columns in 1 data point or row	Sample Content
1. Vertex 1	breakingnewsg
2. Vertex 2	who
3. Relationship	Tweet, mentions, or replies-to
4. Tweet	Ebola virus: 44 new cases and 21 deaths in Sierra Leone, Liberia and Guinea between July 6-8—@WHO http://t.co/QW5vctdvmP
5. Domains in tweet	breakingnews.com
6. Hashtags in tweet	none
7. Tweet date (UTC)	2014/7/11
8. Followed (by vertex 1)	4
9. Followers (of vertex 1)	1,968
10. No. of Tweets (by vertex 1)	33,971
11. Time Zone	Central Time (US & Canada)

Table 2 Data Dimensions Subjected to TDA Analysis

Dimension	Explanation
Relationship	Numerically coded as 1, 2, 3, representing a tweet, mentions, and a replies-to relationship, respectively. (this is the relationship between vertex 1 and vertex 2)
Domains	Numerically coded as 1 or 2. (answers the question as to whether domains are included in the tweets): YES=1, NO=2
Tweets	The number of tweets vertex 1 has generated.
Followed	The number of twitter accounts vertex 1 follows.
Followers	The number of followers vertex 1 has.
Hashtags	Numerically coded as 1 or 2. (answers the question as to whether hashtags are included in the tweets): YES=1, NO=2
Date of Tweet (UTC) in months	Numerically coded according to the numerical order of the months the tweets were posted. 6 for June, and 7 for July etc.
Tweeted term (keyword)	Numerically coded as 1 or 2. (answers the question as to whether the keyword is included in the tweet): YES=1, NO=2

to map the topological network. One data point or row contains 8 dimensions. Non-numerical dimensions such as the relationship between the vertices, the use or non-use of domains and hashtags, the date in months when the tweets were posted, and the inclusion or non-inclusion of the searched keyword “Ebola” in the tweets, were coded as integers. Since a visual approach is taken to extract insights from these complex data, different topological network structures have to be mapped using different metrics, and then visually evaluated. A metric represents similarity or distance between data points or rows. Here, the Hamming, Jaccard, Manhattan

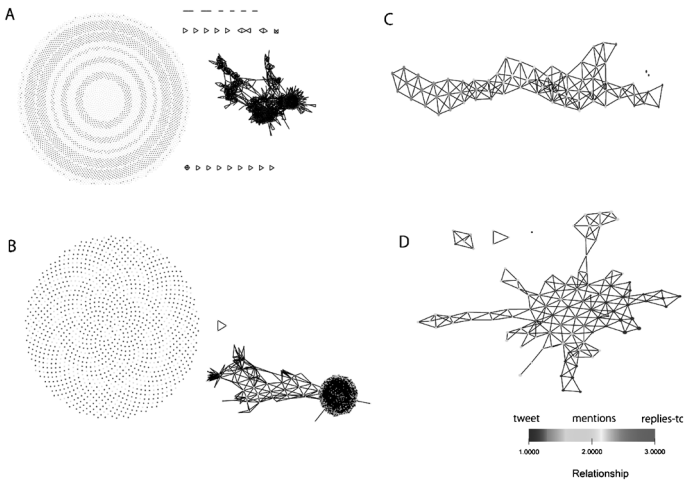

Fig. 2　Network Structures Mapped with Different Metrics
The network structures shown above were plotted using A) Hamming, B) Jaccard, C) Manhattan (L1), and D) Variance Normalized Euclidean. These were plotted with the same neighborhood lenses (resolution: 15, gain: 2). Nodes are colored according to their relationship. Blue, green, and red colors represent tweet, mentions, and replies-to relationships, respectively.

(L1), and the Variance Normalized Euclidean metrics were tested in conjunction with neighborhood lenses. The neighborhood lenses based on k-nearest neighbors were used to emphasize the metric structure of the data. The resolution and gain of these lenses were adjusted to permit the visualization of a meaningful structure. In other words, these adjustments affect the number of clusters. Here, the neighborhood lenses were set at a resolution of 15 and a gain of 2. The choice of distance metric and lenses as well their resolution and gain will affect the layout and cluster size. **Figure 2** shows the network structures mapped with different metrics. One can see that the network structures in A and B, are rather complex with a large number of unconnected nodes, a group of nodes having complex connections, and a few subgroups. Obviously, it is difficult to make sense out of these two network structures. In C, there is a simpler network structure with one connected component and two singletons with no edges or connections to other nodes. Here, there are no subgroups or distinct flares. On the other hand, D clearly has more distinct flares and subgroups than C. However, this is only a visual comparison, and thus the metric that suits the data better needs to be determined.

Table 3 shows the mean and standard deviation values of the quantitative data. The standard deviation values of these dimensions vary considerably, and their means are very much different. Both the Hamming and the Jaccard metrics are inappropriate based on the nature of the data itself. On the other hand, the Manhattan (L1) metric can be applied. However, this metric is best suited when considering that many small differences are as significant as one large difference. Thus, the variance normalized Euclidean, which is best suited for data that

Table 3 Mean and Standard Deviation Values of Quantitative Data Dimensions Subjected to TDA Analysis

Dimension	Mean	Standard Deviation
Tweets	42,247.684	87,389.090
Followed	1,177.711	5,842.321
Followers	8,094.799	137,983.786

contain heterogeneous scale variables, is chosen. This metric finds the mean and standard deviation of each dimension and rescales the value of the coordinate by subtracting the mean and dividing by its corresponding standard deviation.

A data lens based on relationship (tweet, mentions, or replies-to) was added to further see distinct structures and explore similarities or differences between these relationships. The topological network of Ebola tweets shown in **Figure 3** was created from **Figure 2** D, with the application of a relationship-based data lens. Thus, nodes are largely grouped or clustered based on the relationship criteria. As mentioned earlier, nodes with common data points or similarities are connected by edges. Please note that these edges do not represent a "*who-mentions-who*" or "*who-replies to-whom*" relationship as illustrated in **Figure 1**. The resulting topological network can be described as having a tweet main group as well as a tweet subgroup, a mentions main group as well as two mentions subgroups, one replies-to main group, and some singletons.

Statistical tests were performed on groups or features in the Ebola tweet network. The Kolmogorov-Smirnoff (KS) scores were used to identify the distinguishing characteristics of distinct groups or subgroups that were visually identified through their shapes. The topological locations of relevant hashtags, URLs of both

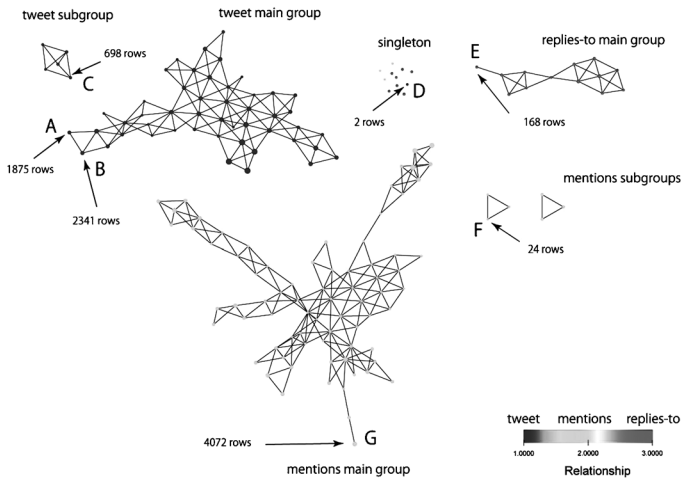

Fig. 3 Topological Network of Ebola Tweets
Here, the network layout was generated with a variance normalized Euclidean metric and neighborhood lenses as with the network structure in **Fig. 2D**, but with an additional lens using the relationship data, which was added to see distinct structures and explore differences between these relationships. Seven nodes (A, B, C, D, E, F, G) were selected from each group to clarify the connections or non-connections between these nodes.

health-related and news organizations, and relevant keywords were identified. Nodes containing vertices that have the highest number of followers as well as those with large centrality values were likewise identified.

3. Results

The topological network of Ebola tweets is shown in **Figure 3**. Here, 74,618 tweets are clustered into 158 nodes with 74,618 rows or data points. A single tweet is treated as one row or a single data point. Nodes represent groups having similar data points. However, one node can represent a single data point or vertex if this has no similarity with other data points. Node size is proportional to the number of data points or vertices contained in a node. Edges connect nodes if they contain common data points. Three main groups representing the basic relationships of the twitter structure, i.e. tweets (48 nodes with 32,572 rows), mentions (75 nodes with 37,477 rows), and replies-to (13 nodes with 1,965 rows) can be readily identified. More diversity can be recognized in the nodes within the mentions main group compared to those in the tweet main group considering the number of nodes in these main groups.

Interestingly, a tweet subgroup and two mentions subgroups distinctly disparate from their main groups, as well as singletons (11 nodes, 371 rows) can be recognized. Although singletons are characterized as either a tweet, a mention, or a replies-to, they have no similarities with their respective relationship groups and as such are disconnected from other nodes. In order to determine why some nodes are connected to other nodes, or for that matter disconnected from other nodes albeit their similar relationship, 7 nodes (A and B in the tweet main group, C in the tweet subgroup, D in the singleton group, E in the replies-to main group, F in one of the two mentions subgroups, and G in the mentions main group) were selected and their similarities were examined by taking their Kolmogorov-Smirnov (KS) scores. The KS score tests the likelihood that two groups being compared have the same distribution of values for a given dimension. The number of rows or data points contained in these nodes is shown in the figure. A comparison of these selected nodes based on the KS scores of the 8 dimensions is presented in **Table 4**. It is clear that nodes A and B have both low KS scores for all 8 dimensions. Large KS values indicate larger differences. Thus, the edge or connection between nodes A and B is confirmed based on the fact that their KS scores are similarly low. For unconnected nodes B (tweet main group) and C (tweet subgroup) which both have a tweet relationship, their KS scores for tweeted term, hashtags, and followers are large i.e. indicating large differences in these dimensions. For that reason, node C appears in a subgroup rather than in the tweet main group.

For unconnected nodes A (tweet main group) and D (singleton) which both have a tweet relationship, they have large differences in almost all of the dimensions (hashtags, followers, tweets, date of tweet, tweeted term, and followed). These two points only have similarities in the use of domains and in their having a tweet relationship. This explains why D is a singleton and unconnected to nodes in the tweet main group or in the tweet subgroup.

For nodes B and E which belong to the tweet main group and the replies-to group, respectively, their main difference primarily lies in their relationship which is obvious.

For nodes F and G which both have a mentions relationship, their main differences as

Table 4 Comparison of Selected Nodes in Figure 3

Compared nodes	Dimension	Kolmogorov-Smirnoff (KS) scores
A (tweet main group) and B (tweet main group)	Tweets	0.1568
	Followed	0.0912
	Followers	0.0841
	Tweeted term (keyword)	8.38×10^{-14}
	Date of Tweet (UTC) in months	8.38×10^{-14}
	Hashtags	8.38×10^{-14}
	Domains	8.38×10^{-14}
	Relationship	8.38×10^{-14}
Compared nodes	**Dimension**	**Kolmogorov-Smirnoff (KS) scores**
B (tweet main group) and C (tweet subgroup)	Tweeted term (keyword)	0.9828
	Hashtags	0.8868
	Followers	0.7051
	Tweets	0.5718
	Date of Tweet (UTC) in months	0.3138
	Followed	0.2575
	Domains	0.0201
	Relationship	4.81×10^{-14}
Compared nodes	**Dimension**	**Kolmogorov-Smirnoff (KS) scores**
A (tweet main group) and D (singleton)	Hashtags	1.000
	Followers	1.000
	Tweets	1.000
	Date of Tweet (UTC) in months	1.000
	Tweeted term (keyword)	1.000
	Followed	0.8454
	Domains	4.62×10^{-14}
	Relationship	4.62×10^{-14}
Compared nodes	**Dimension**	**Kolmogorov-Smirnoff (KS) scores**
B (tweet main group) and E (replies-to group)	Relationship	1.000
	Date of Tweet (UTC) in months	0.4345
	Tweets	0.3006
	Followed	0.2607
	Followers	0.2413
	Hashtags	3.52×10^{-14}
	Domains	3.52×10^{-14}
	Tweeted term (keyword)	3.52×10^{-14}
Compared nodes	**Dimension**	**Kolmogorov-Smirnoff (KS) scores**
F (mentions main group) and G (mentions subgroup)	Domains	1.000
	Hashtags	1.000
	Date of Tweet (UTC) in months	1.000
	Tweets	0.7314
	Followers	0.6692
	Followed	0.6654
	Tweeted term (keyword)	3.96×10^{-14}
	Relationship	3.96×10^{-14}

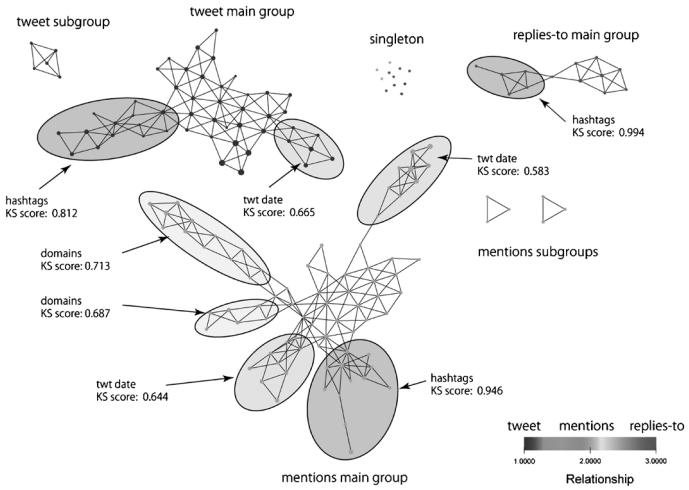

Fig. 4　Topological Network of Ebola Tweets

Node size is proportional to the number of vertices in the node. Nodes are colored by relationship. Blue, green, and red colors represent tweet, mentions, and replies-to relationships, respectively.

Flares which indicate distinguishing characteristics of the topology were visually selected and nodes included in these flares are indicated by colored ellipses. KS (Kolmogorov-Smirnov) scores are used to indicate significant differences in these flares in each of the three main relationship groups.

shown by their KS scores are in the following dimensions: domains, hashtags, date of tweet, tweets, followers and followed. This accounts for why these nodes are not connected by an edge.

Having created the topological network of Ebola tweets, and after examining why some points are connected by edges while some are not, the next step is to look for distinguishing geometric features or structures in the data topology, specifically, flares (long segments) in the main relationship groups. These flares may indicate a subgroup or subgroups (within these main relationship groups) with distinguishing characteristics. **Figure 4** shows the Ebola tweet network structure presented in **Figure 3**, but with the flares in each of the three main relationship groups bounded by colored ellipses for easy identification. As mentioned earlier, the fundamental idea is to take the geometric approach to pattern or shape recognition that allows the discovery of insights in the data. Distinct shapes in the topological network, i.e. flares, and the nodes included in these flares were visually identified and selected. However, visual identification of these features needs to be coupled with the knowledge as to what sets these features apart from the rest of the topological network of each relationship group. Here, the distinguishing characteristics of these flares in terms of the data dimensions mentioned earlier are examined by again using the KS scores. As previously stated, a large KS score indicates a larger difference between compared groups. The hashtags (max. KS : 0.946) and domains or URLs (max. KS : 0.713) dimensions have been identified as the most distinguishing characteristics of these distinct topological features. In the tweet main

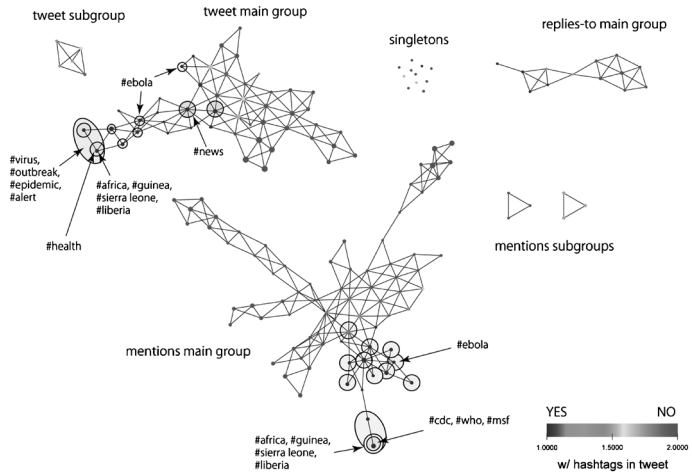

Fig. 5 Topological Location of Relevant Hashtags in Ebola Tweets

Node size is proportional to the number of vertices in the node. Nodes containing vertices with tweets that include hashtags are colored blue; those without hashtags are colored red. Relevant hashtags included in Ebola tweets were identified and mapped onto the topological network. Nodes containing a large number of vertices with tweets having these hashtags are bounded by circles or ellipses for easy identification.

group, the two distinguishing dimensions are the hashtags (KS=0.812) in the left flare, and the tweet date (KS=0.665) in the right flare. In the mentions main group, the distinguishing dimensions are the hashtags (KS=0.946) in the bottom flare, the tweet date (KS=0.583) in the upper right flare, (KS=0.644) in the lower left flare, and domains in the two left uppermost flares, (KS=0.713, and KS=0.687). While in the replies-to main group, the left flare's identifying feature is the hashtags (KS=0.994) dimension.

Hashtags (#) are commonly used on social media sites such as Twitter to identify messages on a specific topic, create a thread of conversations which allows information dissemination beyond one's followers. Pertinent hashtags contained in the Ebola tweet data were examined. **Figure 5** shows the topological location of relevant hashtags in Ebola tweets. The nodes containing vertices with tweets that include hashtags (YES) are colored blue; those that do not (NO) are colored red. Deeper blue nodes contain a high number of YES (hashtag inclusion) nodes, and deeper red nodes contain a high number of NO (hashtag non-inclusion) nodes. Since it is of considerable interest to identify hashtags which are of relevance to the Ebola epidemic, the tweets were searched for Ebola-related hashtags, and then mapped onto the topological network. Event-related (*#ebola*, *#virus*, *#outbreak*, *#epidemic*, *#alert*, *#news*) and location-related (*#africa*, *#sierra leone*, etc.) hashtags, and hashtags of health-related organizations (*#cdc*, *#who*, *#msf*) have been identified. Nodes containing a relatively high number of vertices that use these hashtags are identified in the figure. It is interesting to note that these

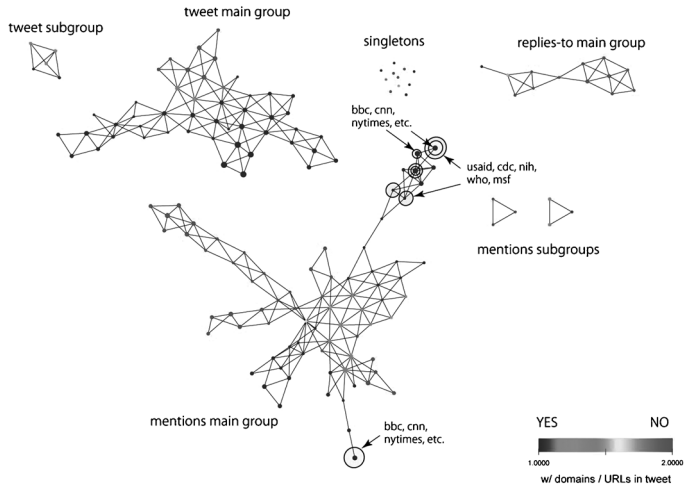

Fig. 6　Topological Locations of URLs of Health-related Organizations and News Organizations Contained in Ebola Tweets

Node size is proportional to the number of vertices in the node. Nodes containing vertices with tweets that include URLs/domains are colored blue; those without are colored red. URLs of health-related organizations and news organizations contained in Ebola tweets were identified and mapped onto the topological network. Nodes containing a large number of vertices having tweets with the indicated URLs are bounded by circles for easy identification.

hashtags appear in large numbers or in greater frequency in the nodes located in the tweet and mentions main groups, at the bottom-left flare and at the bottom-right flare, respectively. Event-related hashtags (*#virus*, *#outbreak*, *#epidemic*, *#alert*) appear in close proximity to or in the same node with the location-related hashtags in the tweet main group. On the other hand, hashtags of health-related organizations (*#cdc*, *#who*, *#msf*) appear in the mentions main group near that of the ebola hashtag (*#ebola*), or together with location-related hashtags. This may indicate the use of multiple but related hashtags in these tweets.

The topological locations of these Ebola-related hashtags coincide with the flares in the tweet main group and in the mentions main group where the KS scores for the hashtag dimensions are high as shown in **Figure 4**. It should be mentioned that the flare in the replies-to main group with a high hashtag KS score contain nodes that include some of the Ebola-related hashtags, but these are not indicated in **Figure 5** since these appear in lesser frequency than the two groups with high hashtag KS scores.

Figure 6 shows the topological locations of some domains or URLs contained in these tweets. Here, the nodes containing vertices with tweets that include domains or URLs (YES) are colored blue; those that do not (NO) are colored red. Deeper blue nodes contain a high number of YES (inclusion of URLs) nodes, and deeper red nodes contain a high number of NO (non-inclusion of URLs) nodes. One may notice that although the top left flare and the center left flare in the mentions main group shown in **Figure 4**

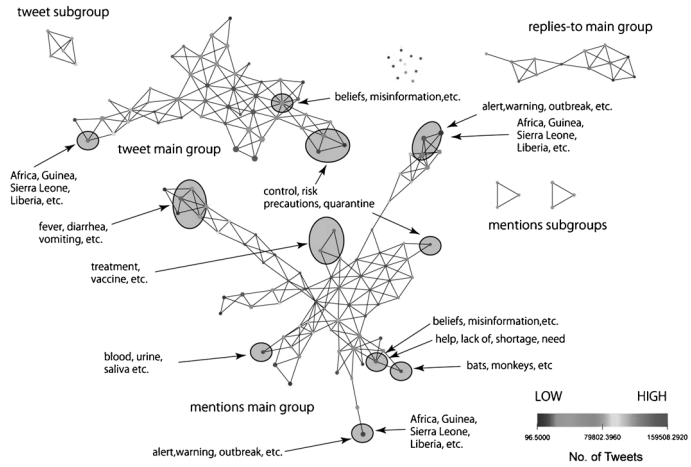

Fig. 7 Topological Locations of Relevant Keywords Contained in Ebola Tweets

Node size is proportional to the number of vertices in the node. Nodes containing vertices with high number of tweets are colored red; those with low number of tweets are colored blue. Relevant keywords contained in Ebola tweets were identified and mapped onto the topological network. Nodes containing a large number of vertices with tweets that include these keywords are bounded by circles or ellipses for easy identification.

indicate high KS scores for the domain dimension, these nodes are colored red which signify non-inclusion of URLs in **Figure 6**. High domain KS score groups do not mean that they have more nodes that include domains or URLs. It means that the domain or URL dimension (inclusion or non-inclusion) are significantly different compared to the rest of the nodes in that group. Domains or URLs of key players in the dissemination of information related to Ebola were identified from tweet data and their locations on the topological network were mapped. Nodes containing a large number of vertices having these URLs are indicated. Domains or URLs of health-related organizations (*usaid*, *cdc, nih*, *who*, *msf*) as well as those of news organizations (*bbc*, *cnn*, *nytimes*) appear only in the mentions main group and not in the other groups. This indicates that these organizations' URLs are mentioned or retweeted by others.

The topological locations of relevant keywords contained in Ebola tweets are shown in **Figure 7**. The nodes are colored by the number of tweets. Nodes with a high number of tweets are colored red; those with low are colored blue. Nodes which contain the highest number of tweets are generally located in the central structure of the main relationship groups. However, attention should be focused not on how many tweets are generated by what vertices, but rather on the number of tweets that contain keywords relevant to the Ebola epidemic. Relevant keywords in these tweets were identified and mapped onto the topological network. These are location-(*Africa*, *Sierre Leone*, *Liberia*, etc.), alert-(*alert*, *warning*, *outbreak*, etc.), symptom-(*fever*, *diarrhea*, etc.), transmission-(*blood*, *saliva*, *bats*, *monkeys*, etc.), and prevention-(*control*,

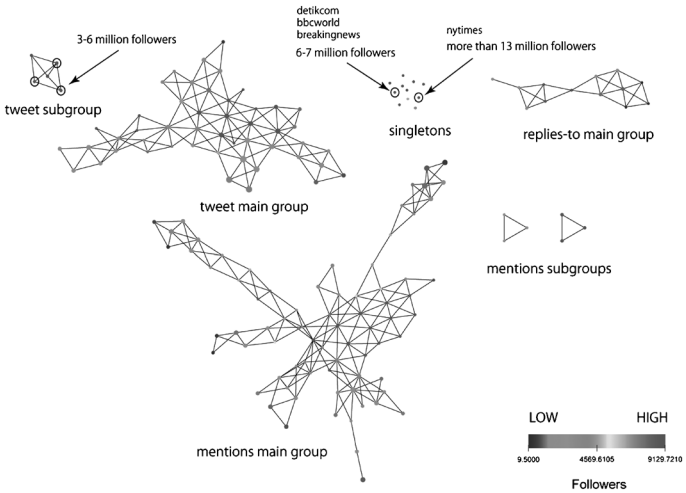

Fig. 8　Topological Locations of Vertices with high Number of Followers

Node size is proportional to the number of vertices in the node. Nodes are colored by the number of followers. Nodes containing vertices with high number of followers are colored red; those with low number of followers are colored blue. Vertices with high number of followers (over 3 million) were identified and mapped onto the topological network.

precautions, *risk*, *vaccine*, *treatment*, *beliefs*, *misinformation*, *need*, *shortage*, etc.) related keywords. Most of the nodes containing a large number of vertices with tweets that include these keywords appear in the flares in both tweet and mentions main groups. The prominence of these keywords indicates a distinguishing characteristic of these flares.

Figure 8 shows the same topological network but colored according to the number of followers. Most of the nodes containing a relatively high number of followers (red nodes in the figure) are shown within the core structure of both the tweet and mentions main groups rather than in the flares. Nodes containing tweets by those vertices with over 3 million followers (mostly news or broadcast networks) are indicated. It is interesting to note that none appear in the tweet main group or in the mentions main group. The *nytimes* (The New York Times) which had the largest number of followers appears as a singleton, i.e. with no connections to other groups.

In social media network analysis, centrality measures are used to identify the most important vertices in the network graph. The in-degree (the number of connections), betweenness (ability to bridge other sub-networks), eigenvector centralities (degree of influence) and page rank (overall importance) were calculated using NodeXL[25]. **Table 5** presents the vertices with high centrality values. **Figure 9** shows their topological locations. Nodes containing a large number of vertices with top centralities are colored red; those with a low number are colored blue. Non-colored nodes do not contain these vertices. Interestingly, these vertices with top centralities, appear in the tweet main group and in its subgroup, in singletons, and in the men-

Table 5 Vertices with High Centrality Values

Centrality Measures	Vertices
In-degree	*who*, *cnni*, *bbcafrica*, *nigerianewsdesk* (news and media website), *natgeo* (National Geographic)
Betweenness	*who*, *bbcafrica*, *cnni*, *nigerianewsdesk*, *nytimes*
Eigenvector	*who*
Page Rank	*who*, *cnni*, *bbcafrica*, *nigerianewsdesk*, *a3noticias* (a Spanish TV station)

tions main and subgroups; none in the replies-to main group. However, nodes containing a large number of vertices with the highest centralities (colored red) appear in the mentions main group, particularly in the flares.

4. Discussion

The TDA approach revealed distinct subgroups and singletons that have no connections with other relationship groups (tweet, mentions, replies-to) although these distinctly separate subgroups or singletons are by themselves a tweet, a mention, or a replies-to. This would not have been possible with openly-available social media networking visualization tools. There is more diversity among Ebola tweets in the mentions main group as compared to those in the main tweet group as indicated by the larger number of nodes (one node contains several data points with similarities) in the mentions main group. Distinguishers of subgroups or flares have been identified with KS scores. This allows us to determine which characteristics de-

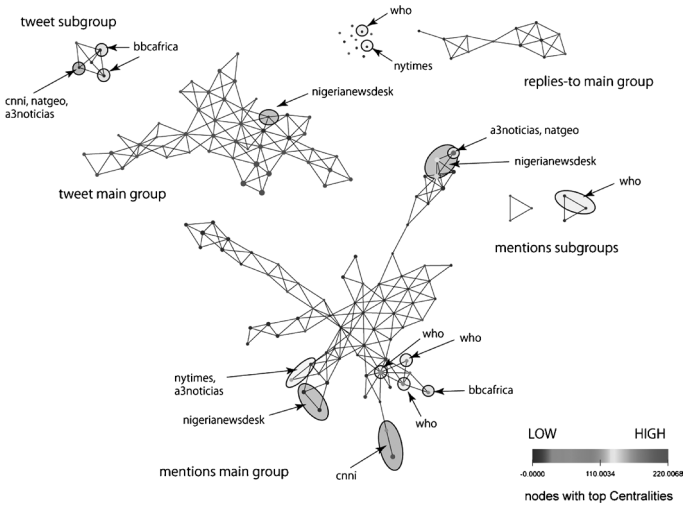

Fig. 9 Topological Location of Vertices with Top Centralities
Node size is proportional to the number of vertices in the node. Nodes containing a large number of vertices (this includes both vertex 1: source of tweet, and vertex 2: destination to which vertex 1 points to) having high centrality measures are shown in red, those with a low number of vertices are shown in blue. Black (non-colored) nodes represent those that do not contain these vertices. Vertices with top centralities were identified and mapped onto the topological network. Circles or ellipses are drawn around these nodes for easy identification.

fine their unique topological features. In this case, the use of hashtags and domains or URLs in tweets has been identified as the most distinguishing characteristics. Since URLs and hashtags have strong relationships with retweetability; and considering that retweeting is a key mechanism for information diffusion in Twitter[26], the presence or absence of hashtags as well as URLs of health- and news-related organizations in Ebola tweets is thus significant. The inclusion of news organizations' URLs also supports the finding in a paper by Boyd et al. that breaking news tends to be retweeted in the form of links to articles in media sources[27].

The inclusion of relevant keywords (location, alert, symptoms, transmission, and prevention) in these Ebola tweets is indicative of the dissemination of information through social media as well as the problems encountered in controlling the outbreak, such as local beliefs and misinformation[28]. By pin-pointing the topological locations of these keywords, the source or sources of misinformation can be identified by examining which subgroup they belong. Public health practitioners can thus formulate strategies to correct misinformation by targeting these sources.

The influence a tweet has on social media is associated with the number of followers an individual or organization has or with centrality metrics[29]. The visual identification of these influencers is thus significant. The cooperation of these influencers in the effort to provide correct and timely health or medical related information can thus be enlisted.

Since most analytical problems, especially those of social or political phenomena are often caused by complex interactions of multiple factors and not by a single cause as argued by Brandes et. al.[30], a visualization tool that can simultaneously display multiple factors is thus necessary.

Here, the TDA approach utilizing Ayasdi Core has made it possible not only to visually render a large, unstructured and highly complex data, but also to present multiple factors (tweet relationships, the number of vertices in a node, the number of followers, etc.) within the topological structure, as well as to identify shapes or structures which is critical in discovering significant characteristics hidden within the data. The locations of vertices with high centralities, relevant hashtags, URLs, and keywords in the flares reinforce the significance of identifying unique structures in the topological network.

Medical informatics has been defined, in a paper by Greenes and Shortliffe[31], as the field that deals with cognitive as well as information processing, and communication tasks of medical practice, education, and research, including information science and technology. According to Haux[32], future research fields of medical informatics may range from interactivity with automated data capture and storage to living labs utilizing data analysis methodologies. A white paper on the secondary use of health data by Safran et. al.[33] emphasizes the expanding volume of health-related data and the growing accessibility of these data. These suggest the importance of managing and analyzing increasingly complex and voluminous data in the practice of medical and health informatics.

5. Conclusion

With TDA, the topological network of Ebola tweets beyond the "*who-mentions-who*" or "*who-replies to-whom*" relationship was visualized and distinct groups and subgroups and their distinguishers were identified. This would not have been possible with social networking analysis

tools. Although this study focused on data mined from social media specifically Twitter, the TDA approach demonstrated here can be used to visualize and analyze large and complex data that health and medical informatics practitioners may need to deal with in the current era of big data. In future work, we intend to apply the same TDA approach on other public health-event related tweets and examine whether similar structure or characteristics can be extracted.

Acknowledgements

The author thanks D. Ramanan of Ayasdi Inc. for valuable comments and suggestions.

References

1) Centers for Disease Control and Prevention. The road to zero-CDC' s response to the West African Ebola epidemic. [http://www.cdc.gov/about/pdf/ebola/ebola-photobook-070915.pdf (cited 2016-jan 12)]
2) Yusuf I, Yahaya S, Qabli S. Role of media in portraying Ebola in and outside Africa. *Journal of Tropical Diseases & Public Health* 2015 ; **3** : 152.
3) Odlum M, Yoon S. What can we learn about the Ebola outbreak from tweets?. *American Journal of Infection Control* 2015 ; **43**, 6 : 563-571.
4) Lancet T. The medium and the message of Ebola. *The Lancet* 2014 ; **384**, 9955 : 1641.
5) Ebola virus: New case emerges in Sierra Leone. [http://www.bbc.com/news/world-africa-35320363 (cited 2016-Jan-15)].
6) Scanfeld D, Scanfeld V, Larson EL. Dissemination of health information through social networks : Twitter and antibiotics. *American Journal of Infection Control* 2010 ; **38**, 3 : 182-188.
7) Hawn C. Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care. *Health affairs* 2009 ; **28**, 2 : 361-368.
8) Gabarron E, Makhlysheva A, Marco L. Type 1 Diabetes in Twitter: Who All Listen To?. *Studies in Health Technology and Informatics* 2014 ; **216** : 972.
9) Heaivilin N, Gerbert B, Page JE, Gibbs JL. Public health surveillance of dental pain via Twitter. *Journal of Dental Research* 2011 ; **90**, 9 : 1047-1051.
10) Sadilek A, Kautz HA, Silenzio V. Predicting Disease Transmission from Geo-Tagged Micro-Blog Data. In AAAI 2012 Jul 22.
11) Krieck M, Dreesman J, Otrusina L, Denecke K. A new age of public health: Identifying disease outbreaks by analyzing tweets. In Proceedings of Health Web-Science Workshop, ACM Web Science Conference 2011.
12) Lamb A, Paul MJ, Dredze M. Separating Fact from Fear: Tracking Flu Infections on Twitter. In HLT-NAACL 2013 Jun (pp. 789-795).
13) Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big data: astronomical or genomical?. *PLoS Biol* 2015 ; **13**, 7 : e1002195.
14) Shneiderman B, Plaisant C, Hesse BW. Improving healthcare with interactive visualization. *Computer* 2013 ; **46**, 5 : 58-66.
15) Inoue M, Hasegawa S, Suyama A, Kakehashi M. Development of a web-based data visualization system for comprehensible ascertainment of the spatiotemporal extent of infectious diseases. *JJMI* 2013 ; **33**, 1 : 27-32.
16) Moon SP, Liu Y, Entezari S, Pirzadeh A, Pappas A, Pfaff MS. Top Health Trends: An information visualization tool for awareness of local health trends. In Proceedings of the 10th International ISCRAM Conference 2013 (pp. 177-187).
17) Dyar OJ, Castro-Sánchez E, Holmes AH. What makes people talk about antibiotics on social media? A retrospective analysis of Twitter use. *Journal of Antimicrobial Chemotherapy* 2014 ; dku165.
18) Smith MA, Shneiderman B, Milic-Frayling N, et al. Analyzing (social media) networks with NodeXL. In Proceedings of the fourth International Conference on Communities and Technologies 2009 Jun 25 (pp. 255-264). ACM.
19) Tinati R, Carr L, Hall W, Bentwood J. Identifying communicator roles in twitter. In proceedings of the 21st international conference companion on World Wide Web 2012 Apr 16 (pp. 1161-1168). ACM.
20) Yoon S, Elhadad N, Bakken S. A practical approach for content mining of Tweets. American *Journal of Preventive Medicine* 2013 ; **45**, 1 : 122-129.
21) Ahn JW, Taieb-Maimon M, Sopan A, Plaisant C, Shneiderman B. Temporal visualization of social

network dynamics: Prototypes for nation of neighbors. In Social computing, behavioral-cultural modeling and prediction 2011 Jan 1 (pp. 309-316). Springer Berlin Heidelberg.

22) Lum PY, Singh G, Lehman A, et al. Extracting insights from the shape of complex data using topology. *Scientific Reports* 2013 ; 3.

23) Carlsson, G. Topological pattern recognition for point cloud data. *Acta Numerica* 2014 ; **23** : 289-368.

24) Carlsson G. Topology and data. *Bulletin of the American Mathematical Society* 2009 ; **46**, 2 : 255-308.

25) Hansen D, Shneiderman B, Smith MA. Analyzing social media networks with NodeXL: Insights from a connected world. Morgan Kaufmann, 2010.

26) Suh B, Hong L, Pirolli P, Chi EH. Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network. In Social computing (socialcom), 2010 IEEE second international conference on 2010 Aug 20 (pp. 177-184). IEEE.

27) Boyd D, Golder S, Lotan G. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In System Sciences (HICSS), 2010 43rd Hawaii International Conference on 2010 Jan 5 (pp. 1-10). IEEE.

28) Oyeyemi SO, Gabarron E, Wynn R. Ebola, Twitter, and misinformation: a dangerous combination?. *BMJ* 2014 ; **349** : g6178.

29) McNeill AR, Briggs P. Understanding twitter influence in the health domain: A social-psychological contribution. In proceedings of the companion publication of the 23rd international conference on World Wide Web companion 2014 Apr 7 (pp. 488 673-678). International World Wide Web Conferences Steering Committee.

30) Brandes U, Kenis P, Raab J. Explanation through network visualization. *Methodology* 2006 ; **2**, 1 : 16-23.

31) Greenes RA, Shortliffe EH. Medical informatics: an emerging academic discipline and institutional priority. *JAMA* 1990 ; **263**, 8 : 1114-1120.

32) Haux R. Medical informatics: past, present, future. *International Journal of Medical Informatics* 2010 ; **79**, 9 : 599-610.

33) Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. *Journal of the American Medical Informatics Association* 2007 ; **14**, 1 : 1-9.