Beyond the Power of Mere Repetition: Forms of Social Communication on Twitter through the Lens of Information Flows and Its Effect on Topic Evolution

1st Yunwei Zhao
CN-CERT
Beijing, China
zhaoyw@cert.org.cn

2nd Can Wang
Griffith University
Gold Coast, Australia
can.wang@griffith.edu.au

3rd Chi-Hung Chi
CSIRO
Hobart, Australia
chihungchi@gmail.com

4th Willem-Jan van den Heuvel
Tilburg University
Tilburg, Netherlands
wjheuvel@uvt.nl

5th Kwok-Yan Lam
Nanyang Technological University
Singapore, Singapore
kwokyan.lam@ntu.edu.sg

6th Min Shu
CN-CERT
Beijing, China
shumin@cert.org.cn

***Abstract*—Understanding how people interact and exchange messages on social networks is significant for managing online contents and making predictions of future behaviors. Most existing research on the communication characteristics simply focuses on the user involvement. The current work largely neglects the content changes that imply how wide and deep the discussion in a topic goes, and to what degree people set forth their own views with the additional information supplemented. We are highly motivated to propose a theoretical framework to target those issues. In this paper, we define the communication modality constructs, and classify topics based on three dimensions: user involvement, information flow depth, and topic inter-relations, which substantially extend the traditional focus in user interaction analysis. The communication modality constructs comprise of (i) topic dialogicity, (ii) discussion intensiveness, and (iii) discussion extensibility. We introduce a quantitative model based on the topology of information flow graph, and use the information addition as well as the emotion attachment along the path to measure the pattern divergence between topic groups. Our model is empirically validated by using 78 million tweets, and experiments on Twitter demonstrate our contributions.**

***Index Terms*—social network, information flow, content mining**

I. Introduction

The understanding on how people interact with each other is significant for managing online contents, making marketing policies [1], as well as predicting sales and transaction trends [2]. For example, the stock market trend forecast of a company is based on the feedback from social media, like Twitter, after the release of a new product [2]. The existing research, however, focuses on the mere replication of information, in lack of a global view on how information is enriched as information flows. For instance, the interaction features are defined with the addressivity (i.e., @ sign) [3], and quantified by the metrics such as the proportions of mentions and replies [4], [5], and etc. Moreover, Sitaram et al.'s work [6] demonstrated that the resonance of the content with the users of the social network (measured by retweet rate) contributes to trend creation and propagation, more than the factors such as user activeness (e.g. tweet rate) and number of followers.

Therefore, rather than the user involvement and the mere message replication, a finer view on the variety of interactions in the diffused topics is lacking. In particular, the identification of the information addition and emotion attached discussions would give a finer view and a more accurate analysis of the customers' feedbacks for business decision making, e.g. utilizing social media for predicting the stock price trend of the company at the pre-release stage of the product. More specifically, the current work suffers from the following:

- Firstly, Twitter is a noisy environment that may communicate and quickly disperse the ambiguous or even wrong information. The additional information (e.g. "reminder: spring sale 20% off iphone cover ends in two days # price" RT @user3: "love it. I cant wait", "RT" is acronym for retweet) provided in messages plays a critical role to enhance the content clarity. The emotional attachment is another key factor, since the emotions transmitted in messages online are "contagious" [7], [8]. People do not live on emotional islands, but rather, that group members experience moods at discussion, and these moods ripple out along with their later online behaviors. This phenomenon is known as "ripple effect" [9].
- Secondly, the uses of multiple hashtags and the co-existing urls indicate another type of modality: topic extensibility. Topics are not self-exclusive, but may extend and possibly overlap with each other in the course of their evolution. This extension is a signal of event significance, in such case, multiple potential hashtags

Corresponding author: Can Wang (can.wang@griffith.edu.au)

978-1-7281-2009-6/$31.00 ©2019 IEEE paper N-19284.pdf

Personal use is permitted, but republication/distribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

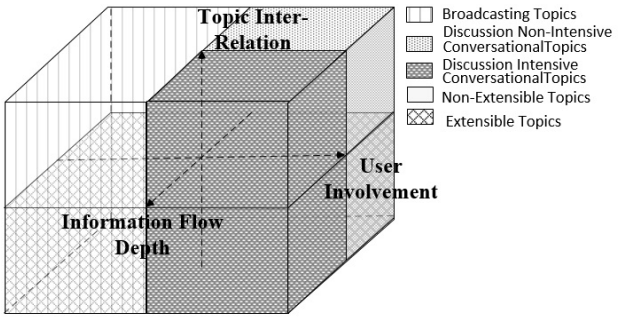

Fig. 1. Focuses of social media communication research.

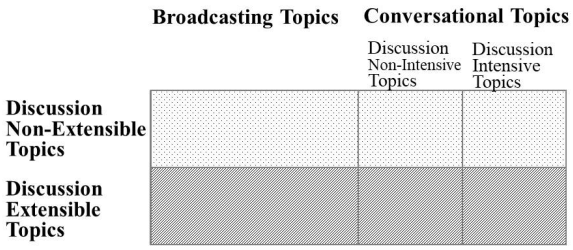

Fig. 2. A Venn Diagram illustration.

are included in one message to ensure that the message reaches the largest range of possible audiences [10]. There are semantic correlations between topics (e.g., "#iphone" and "#engaget" are all related to a broad category "technology"), or shared interests of a certain group of users (e.g. "#free" and "#new" may reveal users key concerns in using them).

Both phenomena can be further analyzed and better understood by mapping to a user-oriented flow topology [11], where the vertices stand for people and the edges represent relationships or communication directions. However, this topology loses the track of content changes during discussions, and mixes multiple discussion threads between users. Rather, we consider introducing an information-oriented flow topology, where the vertices denote the unique content in messages after filtering out the redundant part from its semantic ancestor, and the edges represent the semantic flow directions. It is general practice to resort to a heuristic solution: graph signature (e.g. breadth, depth, in- and out-degree) to describe the graph distribution, as no polynomial time is known for the graph isomorphism problem [12], [13]. However, the generic graph signature does not suffice to induce the various communication modes from the topology itself because (a) it mixes the topology features of the paths that differ in the orthogonal taxonomy dimensions (i.e. user involvement, content change, and the existence of hashtags and urls) identified to distinguish the communication modality constructs; (b) it does not keep track of the variation in the user attention, the emotion attachment, and the information addition along with the content changes. Hence, we propose a neat method based on the graph signatures defined at multiple granularity levels (i.e. vertex, path to topology composition level) that reduces the graph computation complexity and thereby greatly simplifies the model.

In this paper, we propose (i) a theoretical framework of communication modality based on three orthogonal dimensions: user involvement, depth of information flow, and topic inter-relation, as shown in Fig. 1, together with (ii) a quantitative model based on such an information flow graph topology. Our contributions are three-fold: Firstly, we define the communication modality constructs, and classify topics based on them. The second and also our main contribution lies in the construction algorithm of information flow graph that quantitatively depicts how wide and deep a topic diffuses. The third contribution is that we extend the traditional communication focus, and propose to use the information addition and the emotion attachment along the path to measure the pattern divergence between topic groups. Experiments on Twitter demonstrate our contributions.

Below, we specify those three main constructs in our framework:

- We begin with the user involvement dimension and come up with the first construct of communication modality: *Topic Dialogicity*. The topics (i.e. the tweets marked with the same hashtag) are classified into broadcasting and conversational topics, based on whether the discussion involves one person or many people, fall at the left and right octants in Fig. 1. The traditional user-oriented flow do not differentiate between original contents and additional comments, and only focuses on the user-axis, thus falls at the left octants.
- The second construct, *Discussion Intensiveness*, is defined along the depth-axis under the condition "user amount>1" along user-axis. It refers to the degree to which the content in a topic is not simply replicated, but proceeds with changing comments. Therefore, it leads to a finer segmentation within the conversational topics: non-intensive and intensive topics, falling at the front-right and back-right octants.
- The third construct: *Topic Extensibility*, referring to the potential of a topic to provoke a wider discussion, is defined along the "topic interaction" (reflected in the form of co-existing hashtags and urls), is orthogonal to the other two. The topics are classified into extensible and non-extensible groups, see the upper half and lower half octants in Fig. 1.

These topic groups are categorized in a nested way, thus a Venn diagram is provided in Fig. 2, to illustrate (i) the complement relation between broadcasting topics and conversational topics, discussion intensive topics and discussion non-intensive topics, discussion extensible topics and discussion non-extensible topics, and (ii) the intersection relation between discussion extensible topics and discussion non-extensible topics with the other topic groups.

The rest of the paper is organized as follows. In Section II, we review the related literature. The theoretical framework is

presented in Section III. In Section IV, we describe the experiment data collected from Twitter and the empirical approach we adopted in this study. The results and the implications to both academia and practitioners are discussed in Section V. The limitations and future research directions are also described. Finally, Section VI concludes this paper.

II. Related Work

Most research on communication in social media exploits the usage of @sign [3], [5], [14]. There are two strands of research:

Transactional Statistics based: This is the most common approach. Honey and Herring [3] identified 7 functions of @sign, in which "addressivity" (direct a message to others) and "reference" (make reference to others) are most relevant to the interaction analysis, but have different mechanisms. Some work [4], [5] distinguishes between "reply" and "mention" within "addressivity", based on whether it is placed at the beginning of a tweet or somewhere else. "Reference" gives credit to a person (usually the originator of content), and is an indication of how information evolves. The most common practice for "reference" is retweet [15], which is usually applied to characterize interaction features.

Network Structure based: There are two types within this category. The first one focuses purely on the social network structure. Such a network is constructed among bloggers based on the link activities in blogosphere, and among users based on the following relations on Twitter [14]. The second concerns the information exchanges along the network, such as IAD (Interaction-Aware Diffusion) [16], and RAIN (Role-Aware INformation diffusion model) [17]). Neither of them yet differentiates between mere repetition and vivid discussion, and they provide very limited knowledge along with the content change. As for the content variation, the early adopted method is the dynamic topic analysis [18] which requires manual coding as follows: T for "on-topic", P for "parallel shift", E for "explanation", M for "metatalk", and B for "break". Our proposed method expands the "T", "P" and "B" by precisely quantifying how multi-users participate in discussions.

III. Theoretical Framework

A. Problem Definition

Definition 1: (Topic) Given a time interval T, a topic t_i is a sequentially indexed set of all the content $t_i = c_0; c_1; \ldots; c_t$ where c_i is reproduced from hashtag h since it first surfaced in c_0 within the entire observation time interval T.

Denote the set of all the topics as $T = \{t_i\}$, and then to determine topic groups of different communication modalities is to find a partition $P(T)$ on T, where $P(T) = p_j$, p_j=$t|t \in T$ and share the same topological features, i.e. topic dialogicity, discussion intensiveness, etc.

Therefore, we firstly describe how to construct the information-oriented flow graph from the raw time-ordered tweets in III-B. After that, we give the precise definitions of communication modality in Section III-C.

B. Information-oriented Flow Graph

1) Graph Construction: The information flow graph of topic t_i is denoted as $G(t_i) = < V(t_i), E(t_i) >$, where vertices $V(t_i)$ denote the unique content in messages after filtering out the redundant part from its semantic ancestor, and edges $E(t_i)$ denote the information flow direction from ancestor vertices to child vertices. The root v_r is set as the hashtag h of topic t_i.

We note that different diffusion mechanisms may also co-exist in one topic (see Fig. 3a), which leads to the formation of sub-topics in diffusion. Three diffusion mechanisms are considered, namely, hashtag cascade, url cascade, and retweet. Hashtag is a concise and accurate content descriptor (expressed as # followed by a word) used to denote a topic [21]. Within a topic marked with the same hashtag, there may exist several url cascades, i.e. tweets also contain the same hyperlink that directs to external text. There may also exist several sub-topic formed through "Retweet", a reproduction of the content itself within topic t_i [22]. It includes both direct retweet (DT) via clicking retweet button and modified retweet (MT) in the format of "RT @ [the users you give credit to] [original content]". Correspondingly, these different diffusion mechanisms have different information-oriented flow graph construction mechanisms. As expected, the retweet and Url cascade are with well-structured content, making the semantic flow easily detected. Hashtag cascade (here by "hashtag cascade", we refer to those tweets not containing diffusion mechanisms any retweet and urls, but simply with hashtag, see Table I) is with the least structured content, and its construction rule is developed based on text divergence with a predefined threshold. Here, we apply the normalized Jensen-Shannon divergence (JSD) [23], a common way used in the directional semantic flow identification [24].

Table I lists the construction rules for information flow graph with different manners. We use the tweet topic "#iphone 4" as an example (see Fig. 3a), with the constructed graph shown in Fig. 3b, where M_1 and M_2 denotes two original messages that get disseminated, C_1 denotes the added comments to M_1 during its propagation. Note that the direction of information flow in social media context is only one-way since the tweets within a topic are temporally sorted. In hashtag cascade-based information flow detection, it is different from Masucci's work that establishes a directionality index based on the entropy difference of the remained word frequency distribution. To determine the existence of an information flow, we firstly transform the content on two vertices v_1 and v_2 into the word frequency distribution P_{v_1} and P_{v_2} with the word amount n_1 and n_2. The content divergence is then given in Equation (1).

$$d(v_1, v_2) = D(P_{v_1}||P_{v_2}) = \frac{\text{JSD}(P_{v_1}||P_{v_2})}{-\pi_1 ln\pi_1 - \pi_2 ln\pi_2} = \frac{H(\pi_1 P_{v_1} + \pi_2 P_{v_2}) - \pi_1 H(P_{v_1}) - \pi_2 H(P_{v_2})}{-\pi_1 ln\pi_1 - \pi_2 ln\pi_2}, \quad (1)$$

where weights $\pi_i = n_i/(n_1 + n_2)$, $\pi_1 + \pi_2 = 1$, and $H(P_{vi})$ is the Shannon entropy. Values towards 1 indicate more divergence in the two contents. If the content divergence is below the tolerance threshold, it means that the latter

TABLE I
INFORMATION DISSEMINATION WAYS AND INFORMATION FLOW GRAPH CONSTRUCTION

Dissemination ways	**Format**	**Example in topic"#iphone4" in Fig. 3**	**Rule**
Modified Retweet	[comment if applicable] RT the person you give credit to	love it, I can't wait: RT user1 user2: News! #iPhone4 Will Have a Better Camera, Sensitive Back	the content after "RT" is the ancestor of the content after "RT", therefore a directed edge is constructed from v_1 to v_2.
Direct Retweet	an automatic trace marked as "Retweeted by" in the "shared content" span class after clicking the "Retweet" button	Retweeted by username 	if the content has not appeared in the previous vertices, create a new vertex; otherwise, get the vertex vi with the same content, and update the vertex signature (see). That is, $\Delta cc(v_i)$ will increase by 1 and $\Delta covp(v_i)$ will increase by 1 if the user $u \in U(v_i)$, otherwise $\Delta covp(v_i)$ will remain the same.
Url Cascade	[comment if applicable] url	#iphone4 #app new app release. Get it now! http://bit.ly/75yR5B	1. co-existing urls and hashtags are extracted as virtual vertices 2. the remained content is a new vertex, with both the root and the virtual vertices extracted in previous step as ancestors
Hashtag	do not contain any of the above formats, only with hashtags	For unbeatable prices on app #app #iphone4 buy now!	1. calculate the normalized divergence between the current vertex v_c and each of the previous posted tweets with Equation 1 2. get the vertex vminD with the least divergence minD. 3. if minD $\leq$ Divergence Threshold, insert an edge $e = (v_{minD}, v_c)$ into $E(t_i)$; otherwise, insert $e = (v_r, v_c)$ into $E(t_i)$.

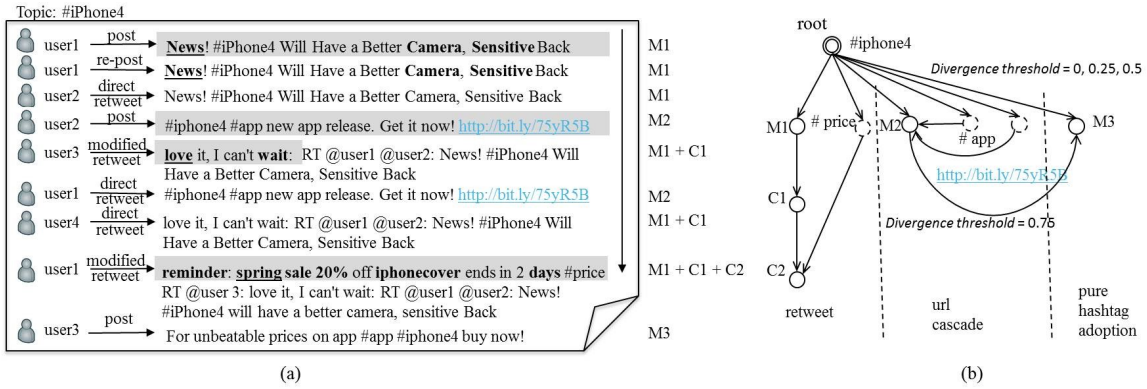

Fig. 3. An illustrating example: (a) original tweets, (b) constructed information flow graph.

post evolves from the previous post. For example, we have $d(M_2, M_3) = 0.69$ in Fig. 3. If the content divergence threshold is set as 0.75, M_2 is the direct ancestor of M_3. Otherwise, the root vertex is the direct ancestor of M_3. The complexity for constrcuting information flow graphs of topic t_i (tweet number is n) disseminated with retweets and urls is $O(n)$, while for hashtag-based is $O(n^2)$.

2) Topology Composition: Topology composition is a collection of paths that are similar to each other with respect to the orthogonal dimensions (i.e. user involvement, depth of information flow graph, and the existence of hashtags and urls) in defining communication modality constructs. Note that even though the communication modality is defined on topics, the orthogonal dimensions that distinguish different communication modalities are actually defined on paths, as a topic may have both monologue parts and dialogue parts, while a path could either be monologue or dialogue, but cannot be both.

The topology composition consists of three levels, as indicated in Fig. 4. The first level consists of intrinsic and extrinsic compositions. The second level: monological and conversational composition is a finer segmentation within intrinsic composition. The third level (also the finest level): discussion non-intensive composition and intensive composition is a finer segmentation within conversational topics.

- *Intrinsic Composition* and *Extensible Composition*. Broadly speaking, the paths is segmented into two parts: intrinsic composition and extensible composition, denoted as $\text{IC}(t_i)$ and $\text{EC}(t_i)$, respectively, based on whether the root vertex of each path is a real or a virtual vertex (e.g. co-existing urls and hashtags).
- *Monologue Composition* and *Dialogue Composition*. These two compositions are segmented within $\text{IC}(t_i)$, with respect denotations given in Equations (2) and (3).
- *Discussion Non-intensive Composition* and *Discusion Intensive Composition*. These two compositions are further segmented within $\text{DC}(t_i)$, which are defined in Equations (4) and (5), respectively.

$$\text{MC}(t_i) = \{p_j(t_i) | \text{uInv}(p_j(t_i)) = 1\}, \quad (2)$$
$$\text{DC}(t_i) = \{p_j(t_i) | \text{uInv}(p_j(t_i)) > 1\}, \quad (3)$$
$$\text{DNIC}(t_i) = \{p_j(t_i) | \text{uInv}(p_j(t_i)) > 1, d(p_j(t_i)) = 1\}, \quad (4)$$
$$\text{DIC}(t_i) = \{p_j(t_i) | \text{uInv}(p_j(t_i)) > 1, d(p_j(t_i)) > 1\}, \quad (5)$$

where $\text{uInv}(p_j(t_i)) = |\cup U(v_k)|$, $U(v_k)$ denotes the set of users on the vertex v_k in the j-th path of topic t_i $p_j(t_i)$, and $d(p_j(t_i))$ denotes the orthogonal dimension "user involvement" and "depth of the information flow" graph.

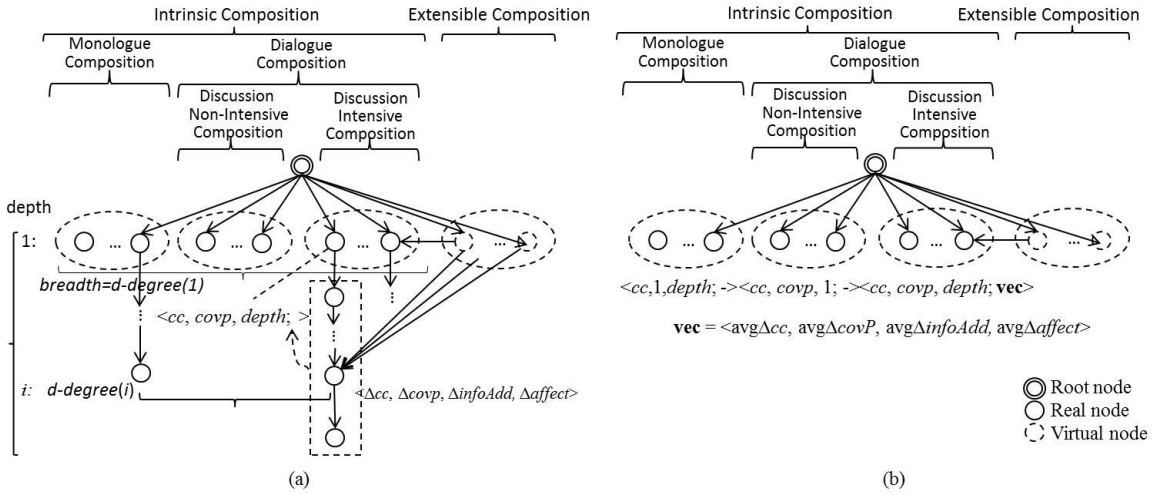

Fig. 4. The topology compositions of a topics information flow graph: (a) detailed view of paths, (b) overview of topology composition.

From those definitions, we can see that there are four types at the finest level: monologue composition, discussion non-intensive dialogue composition, discussion intensive dialogue composition, and extensible composition. Denote all the paths in the topic t_i's graph as $P(t_i) = \{p_j(t_i)\}$. Accordingly, we have $P(t_i) = \text{IC}(t_i) \cup \text{EC}(t_i)$, $\text{IC}(t_i) = \text{MC}(t_i) \cup \text{DC}(t_i)$, $\text{DC}(t_i) = \text{DNIC}(t_i) \cup \text{DIC}(t_i)$.

C. Communication Modality

In this part, we define the communication modality constructs mentioned in Section I and then classify the topics accordingly.

Definition 2: (Topic Dialogicity) A topic t_i is a broadcasting topic if its dialogue composition is an empty set, i.e, $\text{DC}(t_i) = \emptyset$. Otherwise t_i is a conversational topic.

the set of broadcasting topics and conversational topics as $\text{BT} = \{t_i|\text{DC}(t_i) = \emptyset\}$ and $\text{CT} = \{t_i|\text{DC}(t_i) \neq \emptyset\}$, respectively, then we have $T = \text{BT} \cup \text{CT}$, $\text{BT} \cap \text{CT} = \emptyset$.

Definition 3: (Discussion Intensiveness) A conversational topic $t_i \in \text{CT}$ is a discussion non-intensive topic if $\text{DIC}(t_i) = \emptyset$. Otherwise, t_i is a discussion intensive topic.

Denote the set of discussion non-intensive topics and discussion intensive topics as $\text{DNIT} = \{t_i|\text{DIC}(t_i) = \emptyset\}$ and $\text{DIT} = \{t_i|\text{DIC}(t_i) \neq \emptyset\}$, respectively, then we have $\text{CT} = \text{DNIT} \cup \text{DIT}$, $\text{DNIT} \cap \text{DIT} = \emptyset$.

Definition 4: (Discussion Extensibility) A topic t_i is a discussion non-extensible topic if $\text{EC}(t_i) = \emptyset$. Otherwise, t_i is a discussion extensible topic.

Denote the set of discussion non-extensible topics and discussion extensive topics as $\text{NET} = \{t_i|\text{EC}(t_i) = \emptyset\}$ and $\text{ET} = \{t_i|\text{EC}(t_i) \neq \emptyset\}$, respectively, then we have $T = \text{NET} \cup \text{ET}$, $\text{NET} \cap \text{ET} = \emptyset$.

IV. Experiment Settings

A. The Data

The source data from Twitter consists of 78 million tweets posted by 112,044 users during 2010.01-2010.10. The tweets are collected through crawling the followers' and followees' tweets of a random selection of active users, with the number of hashtags more than 1000. The topics in this paper are the tweets marked with the same hashtag, despite of their diffusion mechanisms. They can be distributed through the hashtag cascade, the url cascade, or the retweet. In total, we have 42,475,140 tweets and 488,411 topics under study.

B. Evaluation Metrics

We use the communication count (i.e. *cc* for short) and the coverage of people (i.e. *covp* for short) between each pair of topic groups to evaluate the effectiveness of our work in characterizing multiple communication modality constructs. The former refers to the times that a message is re-sent without changes including retweeting and reposting, and the latter means the number of people that re-send a message without changes. Accordingly, we use the maximal and average values of both metrics to evaluate how topic groups differ with each other: $max(\Delta cc(p_j(t_i))), avg(\Delta cc(p_j(t_i)))$ and $max(\Delta covp(p_j(t_i)))$, $avg(\Delta covp(p_j(t_i)))$, where we have $\Delta cc(p_j(t_i)) = avg(\Delta cc(v_k))$, $\Delta covp(p_j(t_i)) = avg(\Delta covp(v_k))$, $\Delta cc(v_k)$ and $\Delta covp(v_k)$ denote the respect differences in *cc* and *covp* between a child node and a parent node along the path.

For the discussion non-intensive topics and the discussion-intensive topics, we zoom in using "information addition", and "emotion attachment", referring to the average information supplemented, and the average affection attached to the content-changed discussion, respectively. Information addition is measured through the entropy of the remained words after filtering out the stop words. The emotion attachment is measured using *ANEW*(i.e. Affective norms for English Words) [25]. The information addition and emotion attachment of a discussion intensive topic are:

TABLE II
TOPIC DISTRIBUTION W.R.T TOPOLOGY COMPOSITION

Clu. #	Number of Topics	Mean								S.D.
		Monologue composition		Dialogue composition				Extensible composition		
				Discussion non-intensive		Discussion intensive				
		pc^* (%)	d	pc (%)	d	pc (%)	d	pc (%)	d	
1	192157	100	1.04	0	0	0	0.04	55	2.04	0.46
2	184853	100	1.02	0	0	0	0.04	2	0.15	0.73
3	34477	75	1.3	20	1	5	1.09	33	2.23	0.19
4	33688	5	0.12	94	1	1	0.08	37	1.35	0.8
5	26666	14	0.27	0	0.01	86	2.11	31	1.73	0.86
6	16570	0	0	0	0	0	0	14	0.19	0.28

* pc is the path coverage, i.e. the percentage of the number of paths within a composition over all the paths of flow graph, d is the depth of flow graph.

TABLE III
AN OVERVIEW OF TOPIC DISTRIBUTION

Clu.#	Number of Topics	Mean				S.D.
		depth	d-degree			
			max	avg	var	
1	391326	0.99	2.63	2.14	0.43	0.56
2	50345	1.46	1.61	1.3	1.23	0.08
3	25027	1.96	62.42	23.38	24.87	0.11
4	21708	1.63	10.57	5.61	3.74	0.02
5	5	11.2	41168.4	5790.78	11650.86	0.04

$$\text{IA}(t_i) = \frac{\sum_{j=1}^{|\text{DI}(t_i)|} avg(\Delta\text{infoAdd}(p_j(t_i)))}{|\text{DIC}(t_i)|}, \quad (6)$$

$$\text{EA}(t_i) = \frac{\sum_{j=1}^{|\text{DI}(t_i)|} avg(\Delta\text{affect}(p_j(t_i)))}{|\text{DIC}(t_i)|}, \quad (7)$$

where $p_j(t_i) \in \text{DIC}(t_i)$ denotes the j-th path of topic t_i .

Note that $cc(v_k)$, $covp(v_k)$, infoAdd(v_k), and affect(v_k) are defined for v_k, to keep track of the average user attention, information, and affection variations on each vertex along the path. We illustrate them by vertex M_1 and C_1 in Fig. 3.

- $\Delta cc(v_k)$ and $\Delta covp(v_k) = |\cup U(v_k)|$ denote cc and *covp*, respectively. We have $\Delta cc(M_1)$ = 3, $\Delta covp(M_1)$ = 2, $\Delta cc(C_1)$ = 2, $\Delta covp(C_1)$ = 22.
- ΔinfoAdd(v_k) is the entropy of the remained words on vertex v_i after filtering out stop words. The remained words of M_1 and C_1 are: "news camera sensitive" and "love wait" (boldfaced in Fig. 3), with entropy ΔinfoAdd(M_1) = 1.1, ΔinfoAdd(C_1) = 0.69.
- Δaffect(v_k) is the average word affect valence weighted by word frequency. For example, the words with *ANEW* value are: "news" and "love" (underlined in Fig. 3), i.e. Δaffect(M_1) = 0.53, Δaffect(C_1) = 0.87.

V. EMPIRICAL STUDY

A. Topic Information Flow Graph

As indicated, Table II and Table III present the distributions of Twitter topic information flow graph[1] w.r.t the topology composition features and the overall graph features, respectively. The topology composition features are described by (i) path coverage (denoted as pc), the percentage of the paths within a topology composition over all the paths in the information flow graph, and (ii) max depth of all the paths within the topology composition. We apply k-means clustering method on these signatures to get the topic distributions. The optimal number of clusters is decided by the Silhouette [26] model. The overall graph features are described by <depth, max(d-degree), avg(d-degree), var(d-degree)>, where d-degree(i) is the amount of real vertices at distance i [13], max(d-degree) = d-degree(1) = $|IC(t_i)|$, refers to the amount of the real vertices at 1-depth, i.e., the breadth of graph. We have the following observations:

- There exist the clusters with distinctively large discussion non-intensive or intensive compositions (note that for both of these two compositions, there are more than 1 people are involved, that is, these topics are conversational topics) with other compositions holding almost equal, see the 4^{th} and the 5^{th} clusters in Table II. For example, "#nowplaying" has the breadth 92161, the breadth of dialogue compositions is 4183, only occupying 21% of all the paths of topic "#nowplaying", and the maximal path depth is 14.
- Most of topics are star-shaped or nearly star-shaped: the breadth is 2-4 orders of magnitude higher than the depth, and the breadth reduces dramatically as the depth goes up, see d-degree in Table III. Purely chain-shaped topics (large depth and narrow breadth) are trivial, see Clu.#5.

This indicates that (i) the information variation is prevalent with the multi-threads on-going (i.e. each path can be viewed as a thread of sub-topics), and (ii) simply focusing on the user interactions does not differentiate between the mere information replication (i.e. discussion non-intensive composition) and the information variation (i.e. discussion intensive composition). Such a fine clarification w.r.t the user-information flow breadth and depth is necessary and insightful.

B. Pattern Divergence w.r.t Discussion Modality

Table IV reports the distribution of topic groups: (i) broadcasting topics and conversational topics, (ii) discussion non-intensive topics and discussion intensive topics, (iii) discussion

[1]Without loss of generality, the divergence threshold is set as 0.5.

non-extensible topics and discussion extensible topics, corresponding to the three communication modalities: topic dialogicity, discussion intensiveness, and discussion extensibility, respectively. Table V zooms into the distribution of topic group w.r.t the topic dialogicity and the discussion intensiveness within the discussion non-extensible and extensible topics. Most of the topics on Twitter fall within the broadcasting topic groups (70.1%). Discussion non-intensive and intensive topics are evenly distributed within the conversational topics: 10.3% and 10.6%. Discussion non-extensible and extensible topics are evenly distributed: 44% and 56%.

Table VI shows the difference between each pair of topic groups w.r.t the maximum and average communication count and coverage of people per topic. Table VII zooms into the information addition, and the emotion attachment of conversational topics. We can see that there exists a non-negligible divergence between each pair of topic groups. The difference in *cc* and *covp* between each pair of topic groups is about 10 times the smaller value (e.g., the avgCC of broadcasting topics is 1.83, and the avgCC of conversational topics is 11.7). Moreover, we have the following observations:

- Discussion extensible topics have a higher probability of having the larger topic dialogicty and discussion intensiveness, the maximum value of *cc* and *covp* in the discussion extensible topics in Table VI is actually a discussion intensive conversational topic.
- The information addition and the emotion attachment of Discussion intensive topics are notably higher than that of Discussion non-intensive topics, see Table VII.

C. Sensitivity of Content Divergence Threshold w.r.t Discussion Modalities

Fig. 5 shows the changes in the maximal breadth and depth of graph within each communication modality topic group, as the divergence threshold ranges from 0 to 1. We note that:

- The content divergence threshold set at 0 is the harshest criterion, resulting in the greatest breadth and the shortest depth of the information flow graph, see Fig. 5a-5f. The divergence threshold set at 0.75 is the loosest criteria, resulting in the least breadth and the longest depth.
- As the threshold varies, the maximal breadth within a topic group reduces, and the maximal depth within a topic group increases. The largest breadth and the largest depth both occur in the conversational topics, in particular, the discussion intensive and extensible topics, see Fig. 5b, 5d, and 5f, where the corresponding curves are the same. Examples include the previously mentioned "#nowplaying" and "#ff" with 14839, 14225 co-existed hashtags and urls, respectively.
- As the threshold varies, the topic amount of discussion non-extensible and extensible topics stays the same, the topic amount of broadcasting topics and discussion non-intensive topics reduces, and the topic amount of conversational topics and discussion intensive topics increases.

We can see the content divergence threshold mainly influences the conversational topics by affecting the breadth and the depth of the intrinsic components. The curves of maximal breadth and depth intersects when the divergence threshold is $\in [0.25, 0.5]$, which is the interval to choose a reasonable divergence value in the construction of information flow graph.

TABLE IV
DISTRIBUTION OF TOPIC GROUPS

tg_ivs. tg_j	$tg_i(\%)$	$tg_j(\%)$
Broadcasting Topics vs. Conversational Topics	70.1%	20.9%
Non-Intensive Topics vs. Intensive Topics	10.3%	10.6%
Non-Extensible Topics vs. Extensible Topics	44%	56%

TABLE V
DISTRIBUTION OF BROADCASTING TOPICS, CONVERSATIONAL TOPIC WITHIN NON-EXTENSIBLE AND EXTENSIBLE TOPICS

	Broadcasting Topics	Conversational Topics	
		Non-Intensive Topics	Intensive Topics
Non-Extensible Topics	87%	7%	6%
Extensible Topics	73%	13%	14%

TABLE VI
DIVERGENCE W.R.T COMMUNICATION COUNT AND COVERAGE

	Number of Topics	max CC	max Covp	avg CC	avg Covp
Broadcasting Topics	385,955	3090	58	1.83	1.22
Conversational Topics	102,456	89407	14775	11.7	5.2
Non-Intensive Topics	50,630	3083	458	3.45	2.13
Intensive Topics	51,826	89407	14775	20.97	8.9
Non-Extensible Topics	214,789	956	116	0.51	0.42
Extensible Topics	276,932	89407	14775	5.95	2.6

TABLE VII
DIVERGENCE W.R.T INFORMATION AND EMOTION ATTACHMENT

	Number of Topics	max Information Addition	max Emotion Attachment
Non-Intensive Topics	50,630	0.23	0.35
Intensive Topics	51,826	0.68	0.87

VI. CONCLUDING REMARKS

In this paper, we propose a theoretical framework of communication modality composing of three constructs: topic dialogicity, discussion intensiveness, and discussion extensibility, and classify topics accordingly. Experiments on Twitter demonstrate the effectiveness of our work. The first contribution is we extend the traditional user involvement focus, by considering the information flow depth and topic interrelation. The second contribution lies in the graph construction algorithm that precisely quantifies how multi-users participate in discussions, whether the mere replication preserves the original information, or the modification attaches something new.

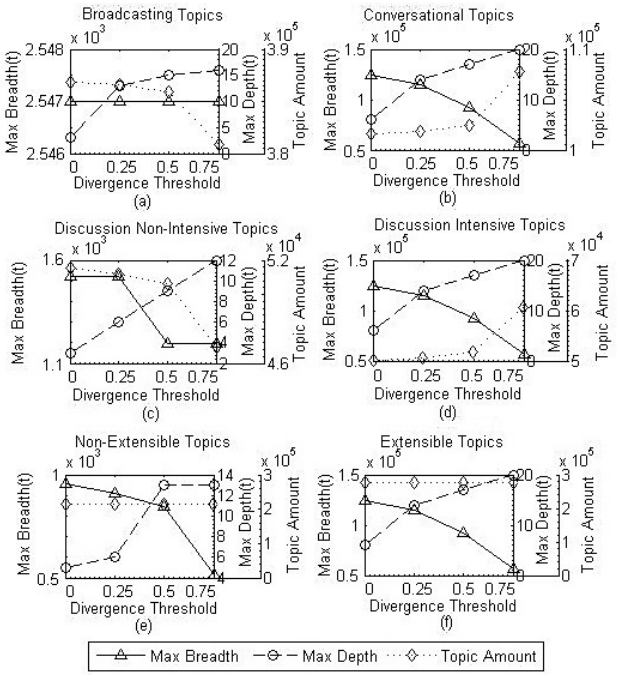

Fig. 5. Sensitivity analysis of content divergence threshold.

Our third contribution is we propose to use the information addition and the emotion attachment along the way to measure the pattern divergence between topic groups. To the best of our knowledge, this is the first framework proposed to target various communication modalities. Besides, our algorithm bears a certain degree of applicability to other time-ordered text-based social networks, e.g. Facebook.

For the B2C mode practitioners, particularly policymakers and marketing departments who intend to get fast feedback on their new products, this study has important implications. As is known, viral marketing occurs largely through CMC interpersonal influence, most commonly through online social networks. The identification of the information addition and emotion attached discussions shed light on the possibility of utilizing social media for business decision making, e.g. predicting the stock price trend of the company at the pre-release stage of the product.

In the future, we may (i) extend the information flow graph based method from the single-media focus to the cross-media setting, in particular, how the information flows between different media; (ii) extend the content divergence measurement with considering latent semantics between tweets with non-overlapping words; (iii) incorporate the communication complexity features into the diffusion model to predict. In addition, the text-based flow direction identification approach proposed for the tweet flow relation specification can also be used to gain a deep understanding of the dynamics of user behavior, such as the followee-follower flow specification. In particular, of all the messages pushed to the followers, it is possible to identify whether there is an influence from a specific followee, and in what form the influence is passed on emotion contagion or information addition. Thereby, users can be segmented into groups for targeted marketing campaigns, for example, recruiting users who are particularly open to sharing information or are emotionally contagious.

REFERENCES

[1] T. Rodrigues, F. Benevenuto, M. Cha, K. P. Gummadi, and V. Almeida, "On word-of-mouth based discovery of the web," IMC, 2011, pp. 381-396.

[2] T. O. Sprenger, A. Tumasjan, P. G. Sandner, I. M. Welpe, "Tweets and trades: The information content of stock microblogs," European Financial Management, vol.20(5), 2013, pp. 926-957.

[3] C. Honeycutt and S. Herring, "Beyond Microblogging: Conversation and Collaboration via Twitter," HICSS, 2009.

[4] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," WWW, 2010, pp. 591-600.

[5] M. Naaman, H. Becker, and L. Gravano, "Hip and trendy: Characterizing emerging trends on Twitter," J. Am. Soc. Inf. Sci. Technol, vol. 62(5), 2011, pp. 902-918.

[6] M. R. Subramani, B. Rajagopalan, "Knowledge-sharing and influence in online social networks via viral marketing," ACM Communications, vol. 46(12), 2003, pp. 300-307.

[7] A. Gruzd, S. Doiron, and P. Mai, "Is happiness contagious online? A case of Twitter and the 2010 Winter Olympics," HICSS, 2011.

[8] Y. Yang, J. Jia, B. Wu, and J. Tang, "Social Role-aware Emotion Contagion in Image Social Networks," AAAI, 2016, pp. 65-71.

[9] S. G. Barsade, "The ripple effect: Emotional contagion and its influence on group behavior," Administrative Science Quarterly, vol.47(4), 2002, pp. 644-675.

[10] R. Abascal-Mena and R. Lema, and F. Sdes, "Detecting sociosemantic communities by applying social network analysis in tweets," Social Network Analysis and Mining, vol. 5(1), 2015, pp. 1-17.

[11] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts, "Who says what to whom on twitter," WWW, 2011.

[12] J. Leskovec, A. Singh and J. Kleinberg, "Patterns of influence in a recommendation network," PAKDD, 2006.

[13] P. C. Wong, H. Foote, G. Chin, P. Mackey, and K. Perrine, "Graph signatures for visual analytics," IEEE Trans. Vis. Comput. Graph, vol. 12(6), 2006, pp. 335-364.

[14] S. A. Macskassy, "On the study of social interactions in Twitter," ICWSM, 2012.

[15] D. Boyd, S. Golder, and G. Lotan, "Tweet, Tweet, Retweet: Conversational aspects of retweeting on Twitter," HICSS, 2010.

[16] Y. Su, X. Zhang, P. S. Yu, W. Hua, X. Zhou, and B. Fang, "Understanding Information Diffusion under Interactions," IJCAI, 2016, pp. 3875-3881.

[17] Y. Yang, J. Tan, C. Leung, Y. Sun, Q. Chen, J. Li, and Q. Yang, "RAIN: Social Role-Aware Information Diffusion," AAAI, 2015, pp. 367-373.

[18] S. Herring, "Dynamic topic analysis of synchronous chat," Symposium on New Research for New Media, 2003.

[19] I. Taxidou and P. M. Fischer, "Online Analysis of Information Diffusion in Twitter," WWW, 2014, pp. 1313-1318.

[20] P. Bogdanov, M. Busch, J. Moehlis, A. K. Singh, B. K. Szymanski, "The social media genome: Modeling individual topic-specific behavior in social media," ASONAM, 2013.

[21] H. Ma, M. Jia, X. Lin and F. Zhuang, "Tag correlation and user social relation based microblog recommendation," IJCNN, 2016, pp. 2424-2430.

[22] N. Xu, G. Chen and W. Mao, "MNRD: A Merged Neural Model For Rumor Detection In Social Media," IJCNN, 2018, pp. 1-7.

[23] A.P. Masucci and V.M. Eguluz and E. Hernndez-Garcia and A. Kalampokis, "Extracting directed information flow networks: An application to genetics and semantics," Phys. Rev. E, vol. 83(2), 2011.

[24] A. P. Masucci, A. Kalampokis, V. M. Eguluz and Emilio Hernndez-Garca, "Wikipedia information flow analysis reveals the scale-free architecture of the semantic space," PLoS ONE, vol. 6(2), 2011, pp. e17333.

[25] A. Bruns and J. E. Burgess, "The use of Twitter hashtags in the formation of ad hoc publics," Proc. 6th European Consortium for Political Research General Conf, 2011.

[26] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," J. Comput. Appl. Math., vol. 20(0), 1987, pp. 53-65.