

Temporal pattern of (re)tweets reveal cascade migration

Ayan Kumar Bhowmick¹, Martin Gueuning^{2,3}, Jean-Charles Delvenne³, Renaud Lambiotte², and Bivas Mitra¹

¹Department of Computer Science and Engineering, IIT Kharagpur, India

²naXys, University of Namur, Belgium

³ICTEAM, Universite Catholique de Louvain, Belgium

ayanb@iitkgp.ac.in¹, martin.gueuning@uclouvain.be², jean-charles.delvenne@uclouvain.be³
renaud.lambiotte@unamur.be², bivas@cse.iitkgp.ernet.in³

Abstract—Twitter has recently become one of the most popular online social networking websites where users can share news and ideas through messages in the form of tweets. As a tweet gets retweeted from user to user, large cascades of information diffusion are formed over the global network. Existing works on cascades have mainly focused on predicting their popularity in terms of size. In this paper, we leverage on the temporal pattern of retweets to model the diffusion dynamics of a cascade. Notably, retweet cascades provide two complementary information: (a) inter-retweet time intervals of retweets, and (b) diffusion of cascade over the underlying follower network. Using datasets from Twitter, we identify two types of cascades based on presence or absence of early peaks in their sequence of inter-retweet intervals. We identify multiple diffusion localities associated with a cascade as it propagates over the network. Our studies reveal the transition of a cascade to a new locality facilitated by *pivotal users* that are highly cascade dependent following saturation of current locality. We propose an analytical model to show co-occurrence of first peaks with cascade migration to a new locality as well as predict locality saturation from inter-retweet intervals. Finally, we validate these claims from empirical data showing co-occurrence of first peaks and migration with good accuracy; we obtain even better accuracy for successfully classifying saturated and non-saturated diffusion localities from inter-retweet intervals.

I. INTRODUCTION

In recent times, Twitter has become one of the most influential micro-blogging systems for spreading and sharing breaking news, personal updates and spontaneous ideas [1], [2]. Twitter provides retweeting facility, through which a tweet simply gets relayed to all the followers of the retweeting user. Long chain of retweets from the original tweet forms the cascade, leading to a rapid information diffusion over the underlying social (follower) network. In state-of-the-art literature, predicting the volume of audience reached by a tweet has been studied extensively in the context of tweet virality. For instance, Cheng et al. [3] studied a sample of large photo reshare cascades on Facebook and predicted future growth of cascades by using temporal and structural features, as well as potential recurrence of the cascade [4]. Similar explorations have been made in [5] and [6]. Weng et al. [7] showed that the initial diversity of a cascade across several network communities is a good predictor for future popularity.

Attempts have been made on modeling and predicting the tweet popularity in terms of the cascade size by studying retweet dynamics using self-exciting point-processes [8], [9].

In Twitter, diffusion of tweets from one user to another occurs mainly via retweeting activities. After each retweet event, a new set of users gets exposed to the content. Importantly, the size of this exposed set is not directly proportional to the follower count of the corresponding retweeter, since the follower set may overlap with the set of users previously exposed to the tweet (via different followees). Notably, this diffusion process in the underlying follower network is not uniform, rather the cascade gets diffused from one locality of the network to another in bursts. In this context, we introduce a *diffusion locality* L^C associated to a cascade C as a connected set of exposed users which *slowly* grows with each retweet activity. Precisely, after each retweet event, if the *set of newly exposed users* is not too large, it is absorbed by the current diffusion locality.

On the other hand, if the set of *newly exposed* users exceeds a threshold, the cascade propagates to a *new* diffusion locality associated to this cascade. Hence, a cascade may propagate to several diffusion localities. The transition of a cascade from one diffusion locality to another plays an important role in the exposition of the tweet to broader audience. Consider two cascades C_i and C_j with the exact same number of retweets. The exposure of cascade C_i will be substantially higher than C_j if C_i migrates to multiple diffusion localities whereas C_j remains confined within a single diffusion locality [7]. The transition of a cascade across multiple localities is facilitated by retweets caused by the *pivotal users*. Identifying the migration of cascade across its diffusion localities is challenging because: (a) The pivotal users responsible for transition are not necessarily the high degree nodes (hubs) of the underlying follower network, and therefore hard to identify even with prior knowledge of the network structure (b) pivotal users are cascade specific and vary widely across different cascades (c) time to explore a diffusion locality and migrate to another one widely varies across diffusion localities and tweet content (d) human (retweet) activity is widely heterogeneous and may be subject to periodic patterns, adding noise in the data.

Retweeting events reflect the human response to a content (tweet), in terms of whether the content is sufficiently stimulating to get repeatedly chosen by multiple users for forwarding. Each retweet in a cascade is timestamped by the time of retweeting the post. In a retweet cascade, the inter-retweet intervals exhibit distinct temporal patterns, which allows to distinguish the cascades in an automated manner. For instance, [10] focused on the distribution of time interval between tweet posts and classified three different user categories: personal (controlled by one person), managed (PR agency controlled) and bot-controlled (automated system). The collective retweeting activities in a cascade varies with the nature of the tweet, leading to diverse (retweet) patterns. In [11], Ghosh et al. leveraged on the retweeting patterns to identify different categories of retweeting activity on Twitter. However, the state-of-the-art literature largely overlooked the potential of temporal retweet patterns to explain cascade transition across multiple diffusion localities. Our study reveals that saturation of a tweet within a diffusion locality makes its content redundant for the population in that locality and subsequently increases the latency between two consecutive retweets. This observation points to the fact that the temporal pattern of a cascade, measured as inter (re)tweet intervals, may work as an indicator for cascade transition across diffusion localities.

The major contribution of this paper is to identify transitions between diffusion localities associated with a cascade only from the inter-retweet intervals. We rely on two empirical datasets containing detailed information of each retweet activity and construct the cascades. We compute the inter-retweet time intervals between two successive posts (section II), and classify the cascades into two types based on the presence or absence of first peaks in the inter-retweet intervals at the early phase of the cascade. We define the diffusion locality of a cascade and illustrate the transition of cascade across multiple localities (section III). We propose an analytical model which explains the co-occurrence between first peaks in the inter-retweet intervals and migration of the cascade to a new diffusion locality. Precisely, the rise-and-fall pattern of the first peak exhibits the saturation (rise) of a tweet in the current locality, followed by migration to a new locality (fall) (section IV). Next we validate the model with empirical observations where we confirm a co-occurrence between the peaks and cascade migration. We show that, at the peak, the post gets saturated in its current diffusion locality, resulting in a steep rise in inter-retweet intervals. Once the post migrates to a new diffusion locality, frequent retweet activity resumes, resulting in a fall in the subsequent inter-retweet intervals. We demonstrate the elegance of the model which successfully classifies the saturated and non-saturated diffusion localities from sequence of inter-retweet intervals obtaining best accuracy of 89% and 94% respectively (section V).

II. DATASET

We collected the following two publicly available tweet datasets [12] connected to the Arab-Spring Movement in

2011 - (i) Algeria Dataset and (ii) Egypt dataset, which are collection of tweets (tweet-ids) and users who posted them during the movement. We crawled the tweet content, user profile and the corresponding follower network for three months. A retweeted tweet provides a link to the original tweet, which allows us to identify all the events belonging to the same cascade. Salient features of the datasets are summarized in Table I. Even though these datasets are relatively small, it has the typical advantage of providing, at the same time, information of cascades and about the underlying social (follower) network.

TABLE I
DETAILS OF ALGERIAN AND EGYPTIAN DATASETS

Dataset	#Tweets	#Retweets	#Cascades	#ActiveUsers	Maximum size of cascade
Algeria	65268	17269	5730	8814	980
Egypt	671417	188090	67539	13882	432

Let the i^{th} retweet in a cascade C of size n denoted by r_i^C , be posted by the user u_i^C at time t_i^C . Then the ordered list of users based on timestamp of posts for cascade C can be represented as $U^C = u_0^C, u_1^C, \dots, u_n^C$ and the corresponding series of retweet timestamps become $(t_1^C, t_2^C, \dots, t_n^C)$. Here u_0^C is the user who posted the original tweet (seed of the cascade). Given the time series $(t_1^C, t_2^C, \dots, t_n^C)$ of a cascade C , we define the sequence of its inter-retweet time intervals T^C as the time interval between occurrence of two consecutive retweets in that cascade. Formally, we denote the i^{th} inter-retweet time interval T_i^C as the interval between posting of the retweets r_i^C and r_{i+1}^C , hence $T_i^C = t_{i+1}^C - t_i^C$. In this paper, we preprocess the data to filter out the cascades of size < 10 (< 30 for Egypt) and obtaining a total of 465 (for Egypt 399) cascades.

III. RETWEET INTERVALS AND DIFFUSION DYNAMICS

In this section, we empirically study the inter-retweet intervals and classify two types of cascades based on the presence of peaks. Side by side, we investigate the cascade diffusion process and introduce the concept of diffusion locality and its saturation in this context.

A. Evolution of inter-retweet intervals

Study of the inter-retweet intervals $T^C = (T_1^C, T_2^C, \dots, T_{n-1}^C)$ of a cascade C reveals the appearance of *peaks* in T^C (see Fig. 1(a) for two different cascades). We define a peak in T^C as a large inter-retweet interval between two consecutive retweet events, indicating a slowdown of the process of cascade diffusion. In the following, we describe a simple methodology to identify the peaks from the series of inter-retweet intervals T^C .

Peak detection in inter-retweet intervals: We apply a simple outlier detection technique [13] on the distribution (lognormal) obtained from the sequence of inter-retweet intervals T^C to detect peaks for cascade C . Let μ_{T^C} and σ_{T^C} denote the mean and standard deviation of the distribution obtained from the sequence T^C . We classify an interval T_i^C as a peak if

$T_i^C > \mu_{TC} + 2 * \sigma_{TC}$. The usage of this threshold ensures the handling of noise in interval data when detecting peaks since the variance of the distribution is large.

Close observation reveals that peaks in the sequence of inter-retweet intervals T^C of a cascade C may appear (a) at the intermediate phase of the diffusion process of a cascade (early peak) and (b) at the last few retweets towards the late phase of almost all the cascades (late peak). Depending on the phase of occurrence of the *first peak* in T^C , we classify cascades into the following two types, as illustrated in figure 1(b).

Type I Cascades: Characterized by the occurrence of its first peak at the intermediate phase (early peak) of their inter-retweet time series T^C , much before the occurrence of its last retweets (figure 1(a)).

Type II Cascades: Characterized by the absence of the peak at the intermediate phase of T^C . The first peak occurs when the last few retweet events take place (late peak) towards the end of the cascade diffusion (inset of figure 1(a)).

Around 80% of the cascades in the Algeria dataset are classified as Type I while 20% of cascades are of Type II. For Egypt dataset, only 61% of cascades are of Type I while 39% of cascades are of Type II. We can discard circadian or any other underlying periodic effect as cause of these peaks since their durations are heavily skewed.

B. Cascade diffusion and saturation of diffusion locality

Next we turn our attention to the diffusion of tweets over the underlying follower network. Let us define a user u in the network as an exposed user with respect to a cascade C if u is a follower of atleast one retweeting user in cascade C . The total set of exposed users for cascade C after all the retweets is denoted as S^C . Let S_i^C denote the set of exposed users after the first i retweets. We compute $P_i^C = \frac{|S_i^C|}{|S^C|}$ as the fraction of users in S^C who have been exposed to cascade C after the i^{th} retweet; thus P_i^C corresponds to the cumulative set of exposed users upto the i^{th} retweet. We plot P_i^C after each retweet i ($1 \leq i \leq n$) in Fig. 1(c) (for a sample cascade C).

Interestingly, Fig 1(c) shows that there is a sudden rise in P_i^C compared to P_{i-1}^C at some retweet i , which indicates a sudden diffusion or exposure of the content to a new population at the i^{th} retweet. We call this sudden rise in the fraction of exposed users P_i^C for a cascade C as a *flush* and the corresponding user u_i^C who posts the i^{th} retweet as a *pivotal user* for the cascade C . Otherwise, we denote the rise in the exposed population as an *accretion*, where the tweet content gets newly exposed to a small population after retweet i .

Detecting flush in a cascade: To identify flushes in a cascade C , we apply the simple outlier detection technique [13] to the sequence of change in the fraction of exposed users. We plot the distribution of the number of flushes in all Type I cascades in inset of figure 1(c); we find that around 43% of such cascades have only a single flush and most Type I cascades (79%) have upto three flushes in the Algeria dataset.

Flush *vis-a-vis* accretion represent the instantaneous property of a cascade, which gets manifested as an outcome of

a single retweet. Collectively, we define diffusion locality¹ L^C associated to a cascade C as a connected set of exposed users *incrementally* growing with each retweet activity. After each retweet i , the locality of the current retweeter may *incrementally* grow by incorporating the set of newly exposed users. Under an *accretion*, the newly exposed population after retweet i will be absorbed within the current locality $L_{current}^C$. Otherwise, in case of *flush* effect, this absorption process fails and a new locality L_{new}^C is discovered, to which the cascade migrates. Thus each new locality is discovered through a *pivotal user* in the network, with a very high non-exposed follower count. A flush in the sequence P^C causes the cascade C to migrate to a new diffusion locality. We define the initial locality L_0^C of a cascade C as the one created by the seed user u_0^C and her followers.

Initially, a newly discovered diffusion locality remains in the *unsaturated* state, when only a few members retweeted the post. However, once exposed to a new content, the members of the locality may start retweeting the post based on their own interest. Finally, we consider that a diffusion locality has *saturated* when almost every retweet activity have occurred from its members.

IV. ANALYTICAL MODEL FOR CASCADE DIFFUSION ACROSS DIFFUSION LOCALITIES

In this section, we propose a simple model of diffusion allowing us to explain the following empirical observations: (a) the presence of a typical early peak in inter-retweet interval for cascades of Type I and the absence of such peaks before last phase for cascades of Type II (b) the co-occurrence of peaks and the saturation of the current diffusion locality followed by the migration into a new locality. Hence this model successfully explains the existence of both Type I and II cascades and their context of emergence in the light of locality saturation. This section is organized in two parts: (i) first, we focus on the cascade diffusion in a single locality, where the model exhibits only a sharp rise in inter-retweet intervals. (ii) second, we extend the model to a more realistic scenario, considering cascade diffusion in multiple localities through pivotal users.

A. Modeling cascade diffusion within a single diffusion locality

The model considers the posting of a tweet from a seed node. As retweets take place, the initial diffusion locality of the seed becomes saturated, until a retweet reaches a new locality. Our calculations focus on the time evolution of the inter-retweet intervals generated by this process. A locality of size ν is approximated by a central node, the seed, with ν neighbors.

Let X_i be the retweet time instance of a user i and f be the corresponding probability distribution between two consecutive retweets of a user². The sorted sample $(X_{(1)}, \dots, X_{(\nu)})$

¹The terms diffusion locality and locality have been used interchangeably in this paper.

² Here, we assume that the time of retweets of the neighbors of seed node are all conditionally independent and drawn from the same distribution f .

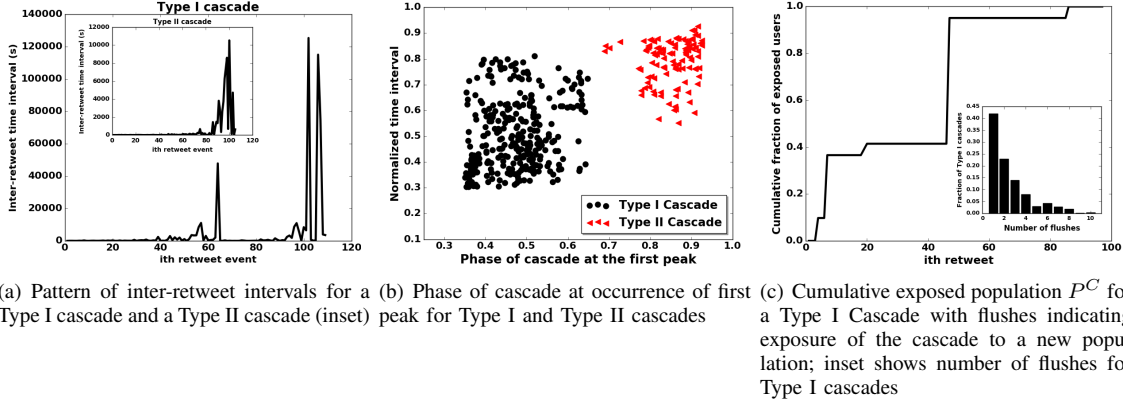


Fig. 1. 2 types of cascades based on phase of occurrence of first peaks (figures 1(a) and 1(b)); cumulative exposed population for a Type I cascade (figure 1(c))

is the observed time series of the cascade. The k^{th} smallest value, $X_{(k)}$, happens at time t with density $f(t)$, provided that exactly $k-1$ events occur before t , and the $\nu-k$ remaining events occur after t . Hence $X_{(k)}$ follows the distribution f_k given by

$$f_k(t) = \frac{\nu!}{(\nu-k)!(k-1)!} f(t) F(t)^{k-1} (1-F(t))^{\nu-k}, \quad (1)$$

where $F(t) \equiv \int_0^t f(\tau) d\tau$ is the cumulative function.

Let E_k be the k^{th} inter-retweet interval, defined as $X_{(k)} - X_{(k-1)}$. The expectation of E_k denoted by $\langle E_k \rangle$ is given by:

$$\langle E_k \rangle = \int_0^{+\infty} t f(t) \frac{\nu!}{(\nu-k+1)!(k-1)!} \cdots F(t)^{k-2} (1-F(t))^{\nu-k} \nu \left[F(t) + \frac{1-k}{\nu} \right] dt \quad (2)$$

In the following, we leverage on $\langle E_k \rangle$ to infer the saturation of a diffusion locality.

Presence of peaks in $\langle E_k \rangle$ for Type I and Type II cascades: In order to explain the occurrence of peaks at different phases of both types of cascades, we analytically solve the model, first considering Markovian process and then a general distribution for f , numerically illustrating the results for various Gamma distributions.

(a) In the simple case in which each retweet is a Markovian process, where $f \sim \text{Exp}(-\lambda)$, the Eq. 2 simplifies to the following closed form equation

$$\langle E_k \rangle = \frac{1}{\lambda} \frac{1}{\nu-k+1}, \quad (3)$$

which is a monotonically increasing function of k . Importantly, when the diffusion locality is not saturated ($k \ll \nu$), we obtain $\langle E_k \rangle \approx 0$, indicating the very low inter-retweet time intervals. On the other hand, when the locality approaches saturation ($k \approx O(\nu)$), we find $\langle E_k \rangle \approx 1/\lambda$, increasing the inter-retweet time intervals.

(b) In situations when f is taken from a general distribution, corresponding to a general renewal process, the Eq. 2 cannot

be explicitly solved. In the saturation phase when almost every member in the locality has retweeted the message ($k \approx O(\nu)$), using a Stirling approximation, one can show that $\langle E_k \rangle$ tends to 0 when k decreases, implying larger values of $\langle E_k \rangle$ for last inter-retweet intervals (high k) of a cascade. This finding is confirmed by numerical simulations in Fig. 2, for a variety of Gamma distributions, which also indicate that the values of last inter-retweet times increase with the variance, while the behavior in the non-saturation phase may vary depending on the system parameters.

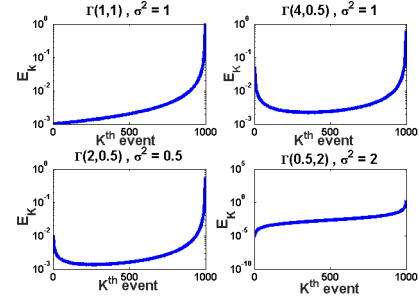


Fig. 2. Evolution of inter-retweet times when activity distribution is $\Gamma(a,b)$; we observe a high increase of the last inter-retweet times for all distributions at the saturation phase

B. Modeling cascade migration across multiple localities

We now consider the interaction between two diffusion localities L_1 and L_2 , which we model as two central nodes H_1 and H_2 respectively connected to each other, and their neighbors. The first central node H_1 triggers a post which gets exposed to its ν_1 neighbors, including H_2 , generating retweet instances (X_1, \dots, X_{ν_1}) . Potentially, when H_2 retweets the post, it leads to an exposure to the new locality formed by H_2 and its ν_2 neighbors, generating retweet instances (Y_1, \dots, Y_{ν_2}) . In the observed (sorted) time series of the cascade $(Z_{(1)}, \dots, Z_{(\nu_1+\nu_2)})$, let the p^{th} retweet correspond to the retweet of H_2 , and the q^{th} to the one of the first neighbor of H_2 who retweets ($q > p$). Similar to E_k , we introduce F_k as the k^{th} inter-retweet interval of the cascade,

defined as $Z_{(k)} - Z_{(k-1)}$. From the definition of q & p , we find $Z_{(k)} = X_{(k)}$ for $k < q$ (diffusion within L_1 , locality of H_1), $Z_{(q)} = X_{(p)} + Y_{(1)}$ (first retweet in L_2 , locality of H_2) and $Z_{(q+1)} = \min(X_{(p)} + Y_{(2)}, X_{(q)})$ (subsequent retweets, in the locality of L_2 or L_1). Hence $F_k = E_k$ for $k < q$, and $F_{q+1} \leq Y_{(2)} - Y_{(1)}$, which attains a low value.

Now there can be two possibilities: (i) If locality L_1 saturates before retweets occur in L_2 ($q \approx O(\nu_1)$), F_{q-1} observes a rise (since $F_{q-1} = E_{q-1}$ and following section IV-A, E_{q-1} gets a high value), and subsequently $F_{q-1} \gg F_{q+1}$. This results in a rise and sharp fall in the consecutive values of F_k , showing a peak in the time series at events $q-1$ to q (explains Type I peaks). Since q is lower bounded by p , this pattern will also be observed when H_2 retweets after L_1 saturates ($p \approx O(\nu_1)$). (ii) On the contrary, if locality L_2 retweets before L_1 gets saturated ($q \ll O(\nu_1)$), F_{q-1} will get a low value (following section IV-A) and hence the decrease in F_k will be too small for the fall to be observed as both F_{q-1} and F_{q+1} are low. This explains the absence of early peak in Type II cascades.

Depending on the presence of such a second locality L_2 and the value of p , we observe the following three cases, illustrated through numerical simulations in Fig. 3.

(a) Case 1: Cascade starts propagating from H_1 and reaches saturation observed by a peak in E_k ; H_2 does not retweet, therefore the cascade does not reach its associated locality L_2 . Type II cascade is produced.

(b) Case 2: Cascade propagates from H_1 ; H_2 retweets and cascade reaches L_2 at an early phase, when L_1 is not yet saturated ($p \ll O(\nu_1)$). The lack of time-scale separation for retweets in L_1 and L_2 does not allow for the observation of a peak. Type II cascade is produced.

(c) Case 3: Cascade starts from H_1 ; H_2 retweets and cascade reaches L_2 at a late phase, when locality L_1 is already saturated ($p \approx O(\nu_2)$). The consecutive retweet events happen around H_2 , and thus small values of the inter-retweet times in H_2 follow the large values in H_1 which leads to an observation of a peak at an intermediate stage. Type I cascade is produced.

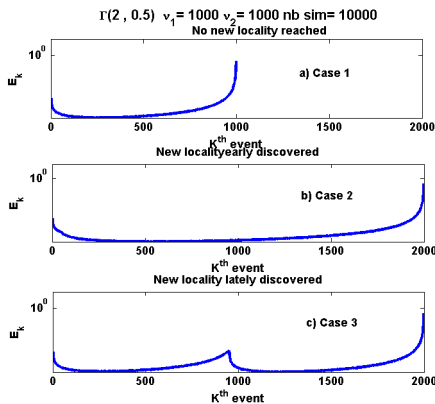


Fig. 3. Migration to a new locality when inter-retweet times follow $\Gamma(a, b)$ distribution. Cascades of Type II are observed when there is no migration or when migration to L_2 happens before saturation of L_1 (case 1 & 2) as in subplots a) and b); cascades of Type I (early peak) are observed for migration to L_2 after saturation of L_1 (case 3) as in subplot c)

In a nutshell, our model demonstrates that (i) inter-retweet intervals $\langle E_k \rangle$ provide signatures of the content saturation in the current diffusion locality, and (ii) peak in the inter-retweet interval (manifested by the rise and fall in E_k) indicates the migration of the cascade to a new diffusion locality after saturation of the current locality.

V. EMPIRICAL VALIDATION

In this section, we empirically validate the claims made in the analytical model, developed in section IV: A) the relationship between inter-retweet intervals and content saturation in diffusion localities, and B) the migration of the cascade to a new diffusion locality following content saturation.

A. Co-occurrence of flush effect and first peaks

First, we exhibit the co-occurrence of first peaks in inter-retweet intervals with the migration of cascades across different localities. In real cascades, migration across diffusion localities gets manifested by the flush effect. In Figure 4(b), we plot the sequence of fraction of exposed users P^C after each retweet event and the corresponding inter-retweet time intervals T^C for a typical Type I cascade C ; we observe that the first peak encountered in T^C is usually located near a flush observed in P^C . The occurrence of the flush may be succeeded by this peak with a short time shift as explained in the model with parameters p and q (section IV-B). We estimate this co-occurrence for all the Type I cascades in our dataset, allowing a maximal shift of three events, and observe the co-occurrence between peak and flush for 82% of cascades in Algeria dataset, as shown in Figure 4(a), and for 69% of such cascades in Egypt dataset. Thus, we empirically validate the claim of the model that the first peak in inter-retweet intervals relates to identifying a new diffusion locality. However, for Type II cascades, we notice the absence of early peaks and the first peaks occurring at the end of the cascade are not accompanied by any flush (see Figure 4(a)).

B. Cascade diffuses to new locality after first peak

In section IV-B, the analytical model showed that retweets of Type I cascades, after the first peak, mainly occurs in the new diffusion locality. In order to validate, we show that after the peak, the retweeters mainly belong to a new diffusion locality.

We define four sets of users for the retweet sequence T^C of cascade C : (1) the set of users \mathcal{P} retweeting around the peak of inter-retweet intervals (mostly the pivotal users), (2) the set $F_{\mathcal{P}}$ of the followers of \mathcal{P} , (3) the set of users \mathcal{B} who retweet before the p^{th} retweet (preceding retweeter of \mathcal{P}) and their followers, and (4) the set of users \mathcal{A} retweeting after p^{th} retweet (succeeding retweeters of \mathcal{P}).

Now we compute the fraction of exposed users f_B^C after the first peak, who already got exposed to the post prior to the peak (i.e., fraction of users in \mathcal{A} that are also in \mathcal{B}); $1 - f_B^C$ denotes the fraction newly exposed to the tweet after the peak. In Fig. 4(c), we exhibit a glimpse of the ratio $r_{new}^p = \frac{1 - f_B^C}{f_B^C}$ for all retweets i of a typical Type I cascade C ; importantly, this

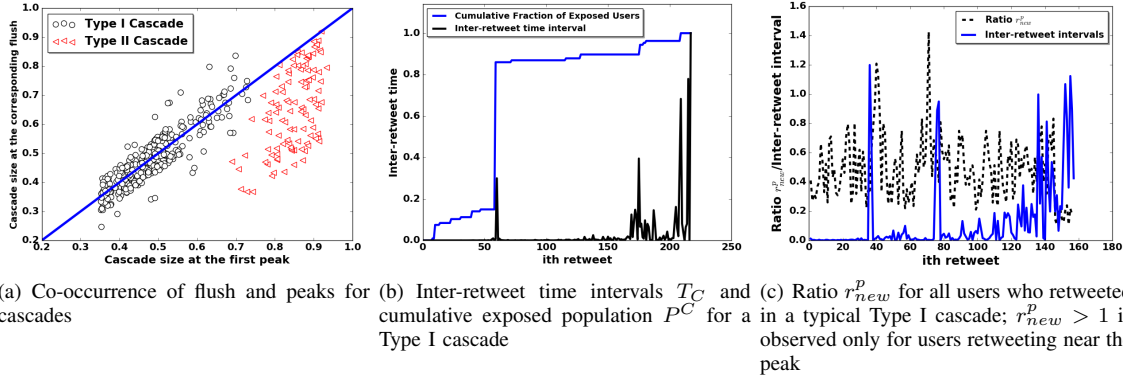


Fig. 4. Relation between first peak and flush effect indicating new locality

ratio becomes greater than 1 only for retweets when a peak occurs in sequence of inter-retweet intervals T^C . In a broader landscape, we observe that the ratio $r_{new}^p > 1$ for 72% of Type I cascades in Algeria dataset; this indicates to the flush effect after the first peak, where the post gets exposed to the new diffusion locality. This result confirms our finding that migration of the cascade to a new locality happens after the first peak where majority of the activity takes place following the peak. On the other hand, r_{new}^p is always < 1 for Type II cascades, indicating the absence of migration of the post to a new locality.

C. Relation between locality saturation and inter-retweet intervals

Finally, we empirically validate the claim made in the analytical model that the inter-retweet time intervals are low ($E_k \approx 0$) when the diffusion locality is not saturated ($k \ll \nu$), and high ($E_k \gg 0$) as it approaches saturation ($k \approx O(\nu)$). For each real cascade, we consider the inter-retweet interval at the first peak as the reference point; less than 20% of this interval can be considered as ‘low’ and more than 80% is considered as ‘high’³. In Egypt (and Algeria) dataset, we observe that when the saturation level ($\frac{k}{\nu}$) is lower than 0.3, for all the cascades, almost 94% (92% for Algeria) of corresponding inter-retweet intervals exhibit low value. On the contrary, when saturation level is greater than 0.8, 89% (82% for Algeria) of the corresponding inter-retweet intervals exhibit high value.

VI. CONCLUSION

This paper puts forward the message that the time sequence of retweet cascades carries unique signature to detect the (a) saturation of a tweet content within a single diffusion locality, and (b) migration of the tweet across multiple localities. We have classified Type I and Type II cascades based on the presence or absence of early peak in the inter-retweet time intervals, obtained from the retweet sequence. Essentially, this peak is the manifestation of latency observed between two

successive retweet events. We have introduced the concept of diffusion locality, which describes the population exposed to the content of a tweet. We have proposed an analytical model which exhibits the co-occurrence of first peaks in inter-retweet intervals with the migration of a cascade to a new diffusion locality facilitated by a pivotal user. This analytical model has been developed in a step by step manner; first we focus on the cascade diffusion within a single locality, next we extend the model considering multiple localities. We have showed that once a tweet content gets saturated within a single locality, latency between two successive retweets observes a steep rise. Subsequently, when the post migrates to a new diffusion locality, frequent retweet activity resumes, resulting in a fall in the inter-retweet intervals. We have validated the model with empirical datasets where we confirm the co-occurrence of first peaks and migration with good accuracy.

REFERENCES

- [1] F. Toriumi, T. Sakaki, K. Shinoda, K. Kazama, S. Kurihara, and I. Noda, “Information sharing on twitter during the 2011 catastrophic earthquake,” pp. 1025–1028, WWW, 2013.
- [2] E. Tonkin, H. D. Pfeiffer, and G. Tourte, “Twitter, information sharing and the london riots?,” pp. 49–57, ASIS&T, 2012.
- [3] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, “Can cascades be predicted?,” pp. 925–936, WWW, 2014.
- [4] J. Cheng, L. A. Adamic, J. M. Kleinberg, and J. Leskovec, “Do cascades recur?,” pp. 671–681, WWW, 2016.
- [5] J. Yang and S. Counts, “Predicting the speed, scale, and range of information diffusion in twitter,” pp. 355–358, ICWSM, 2010.
- [6] S. Pramanik, Q. Wang, M. Danisch, S. Bandi, A. Kumar, J.-L. Guillaume, and B. Mitra, “On the role of mentions on tweet virality,” pp. 204–213, DSAA, 2016.
- [7] L. Weng, F. Menczer, and Y.-Y. Ahn, “Predicting successful memes using network and community structure,” ICWSM, 2014.
- [8] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, “Seismic: A self-exciting point process model for predicting tweet popularity,” pp. 1513–1522, KDD, 2015.
- [9] R. Kobayashi and R. Lambiotte, “Tideh: Time-dependent hawkes process for predicting retweet dynamics,” AAAI, 2016.
- [10] G. Tavares and A. Faisal, “Scaling-laws of human broadcast communication enable distinction between human, corporate and robot twitter users,” vol. 8, no. 7, p. e65774, PloS one, 2013.
- [11] R. Ghosh, T. Surachawala, and K. Lerman, “Entropy-based classification of ‘retweeting’ activity on twitter,” *arXiv preprint arXiv:1106.0346*, 2011.
- [12] A. Bruns, T. Highfield, and J. Burgess, “The arab spring and social media audiences,” vol. 57, no. 7, pp. 871–898, ABS, 2013.
- [13] S. Seo, *A review and comparison of methods for detecting outliers in univariate data sets*. PhD thesis, University of Pittsburgh, 2006.

³Notably, we repeated the experiments with variety of thresholds; nevertheless, the outcomes remain consistent.