2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Uncovering News-Twitter Reciprocity via Interaction Patterns

Yue Ning, Sathappan Muthiah, Ravi Tandon, and Naren Ramakrishnan
Discovery Analytics Center, Department of Computer Science, Virginia Tech
900 N Glebe Rd, Arlington. VA 22203. USA
Email: {yning, sathap1, tandonr, naren}@vt.edu

***Abstract*—In recent years, the amount of information shared (both implicit and explicit) between traditional news media and social media sources like Twitter has grown at a prolific rate. Traditional news media is dependent on social media to help identify emerging developments; social media is dependent on news media to supply information in certain categories. In this paper, we present a principled framework for understanding their symbiotic relationship, with the goal of (1) understanding the type of information flow between news articles and the Twitterverse by classifying it into four states; (2) chaining similar news articles together to form story chains and extracting interaction patterns for each story chain in terms of interaction states of news articles in the story chain, and (3) identifying major interaction patterns by clustering story chains and understanding their differences by identifying main topics of interest within such clusters.**

I. Introduction

Social media sources like Twitter, Facebook, Instagram, Reddit etc. have grown to become an effective part of one's daily life. Twitter has emerged as a powerful medium where people report and comment on everyday happenings. With the proliferation of social media, information shared in traditional media sources like news and blogs is no longer independent of the information in social media; there is implicit information exchange across them. Twitter for instance, tends to break developments rapidly for events that involve mass public involvement such as sporting events and natural disasters [1]; news on the other hand is still the prime source for events related to politics and government.

Traditional and social media sources thus share a symbiotic relationship. In many scenarios, traditional news media is dependent on social media to help spread its news to the masses whereas in other scenarios, social media is dependent on traditional media to supply new information to comment/feed upon. Such interdependencies tend to vary based on the popularity of a topic in social media and also on the geographic location of the topic. In this paper we try to uncover such symbiotic relationships through a principled framework by identification of *interaction patterns* between news and tweets, understanding the differences in such interaction patterns, and imputing such differences corresponding to distinct information topics.

To illustrate an example interaction pattern, Fig. 1 showcases a series of news reports following a fire accident at a nightclub in Santa Maria, Brazil on 27th Jan 2013 as detected by our framework. This figure also depicts trends in Twitter with respect to keywords and actors (persons and organizations) mentioned in news reports. From the Twitter trends, we can see that for certain news reports there is a peak in the corresponding Twitter activity profile *before* its publication time, whereas for some, such peaks happens *after* the news report's publication. This observation suggests that we can use such timing and volume information to capture the direction of information flow between news and Twitter. For instance, in Fig. 1, the newswire breaks the story first. This news was possibly captured by Twitter next as there is a spike in Twitter activity before the second news article. News then immediately follows up and this way both news and Twitter reciprocate. Throughout this *story chain* progression, interactions between news and Twitter activity are clearly evident and mining this interplay and reciprocity is the goal of the framework proposed in this paper. To the best of our knowledge, this is one of the first approaches to do such a study. We next summarize the main contributions of this paper:

- We present an online story chaining algorithm which *chains* related news articles together in a low complexity manner. Our algorithm is based on weighted scores of similarities across news articles for three sets of features: textual features (related to keywords), spatial features (such as locations and geographical coordinates), and actors (such as person(s), and organizations mentioned in the articles).

- We introduce a mechanism to classify the interaction between a news article and Twitter activity around its publication time through four *interaction states*: N (information flow from news article to Twitter), T (information flow from Twitter to news article), B (bi-directional interaction between news and Twitter), and E (empty, or no interaction). This encoding mechanism is applied to all articles in a story chain resulting in a string of interaction states; and the collective string is the *interaction pattern* of a story chain.

- We identify the major source of information for a given story chain based on the interaction states of every news report in the chain and its corresponding quantitative weights. To this end, the interaction patterns of story chains are used to identify distinctive clusters of interactions. Distinct clusters of interaction patterns are further studied to check for clear and explicit dissimilarities in terms of the content reported by the news articles in each cluster. LDA-based topic modeling is used to discern content differences between the different interaction pattern clusters.

II. Related Work

Three categories of related work are briefly discussed here.

Storytelling (or "connecting the dots") as a data mining concept was introduced by Kumar et.al. in [2]. It aims to

ASONAM '15, August 25-28, 2015, Paris, France
 ISBN 978-1-4503-3854-7/15/08 $15.00
DOI: http://dx.doi.org/10.1145/2808797.2809329

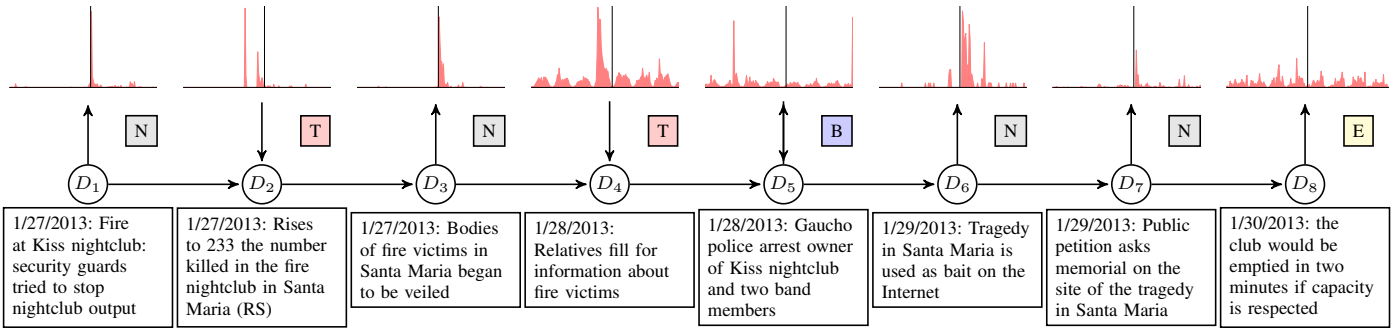


Fig. 1: An example of a story chain from Brazil about a fire accident at a nightclub. For every news report (circles), its corresponding twitter activity profile is shown. The activity profiles are centered around the article publication time and direction of arrow from news reports to the activity profile indicate the direction of information flow and thereby interaction type. Observe the multiplicity of directionalities in this example.

relate given start and end documents by uncovering a series of intermediate documents. This problem has been studied in a variety of contexts such as entity networks [3], social networks [4], cellular networks [5], document collections [6], [7], [8], [9]. Most existing approaches to storytelling [6], [7] use offline data wherein a user must specify the start and end documents of the chain and the algorithm aims to uncover the sequence of relationships between the two endpoints. Shahaf et al. in [8], [9] define concepts of chain coherence, coverage, and connectivity offering insights into the storytelling process. This approach relies on building bipartite word-document or word cluster graphs making it computationally expensive. Leskovec et al. [10] develop a meme-tracking approach for online text and observe a "heartbeat"-like pattern in the handoff between news and blogs.

Twitter's role in event reporting and as a news source is well established. Sakaki et al. [1] used Twitter users as sensors to estimate locations of events such as earthquakes. Chierichetti [11] et al. analyzed tweet streaming to identify important events and the tweet production/consumption patterns around the key events. They observed a robust "heartbeat" phenomenon when key events happen. Ramakrishnan et al. [12] use Twitter with other data sources to forecast protests and civil unrest. Y. Hu et al. [13] present a joint Bayesian model framework called ET-LDA to extract topics covered by the event and the tweets and to perform event segmentation in one unified framework. Jin et al. [14] proposed a topic model which learns topic distributions for two datasets by transferring topical knowledge.

Communication patterns of a social network within itself and with external platforms have been explored with diverse techniques. Hopcroft et al. [15] have studied the reciprocal relationship in a dynamic social network and their findings suggest how individuals' behavior are determined by social structures. There have been studies [16], [17], [18] of the relationship between Twitter and traditional news media and especially how fast and powerful Twitter can be for publishing or discussing live stories. Also, the role of Twitter in news reporting has been explored [18]. Petrovic et al. [17] examine the extent to which news reports and Twitter overlap and whether Twitter often reports faster by manually identifying major news events. Kwak et al. [19] studied the topological characteristics of Twitter as a platform of information sharing. Regarding connecting tweets to news, Sankaranarayanan et al. [20] developed a news processing system called TwitterStand to capture tweets that correspond to late breaking news.

The above efforts chip away at the problem of modeling the interaction between news and social media but only address partially our goals here. They either focus on how news articles can be chained together to study news-news interaction or study about how Twitter can replace news. At the other extreme, while studies such as Petrovic et al. [17] look at the overlap between news and Twitter, these works require significant human involvement. In our framework, we combine temporal dynamic characteristics of tweets and align them to news articles for each story. We define interaction patterns in both quantitative and qualitative ways for story chains and cluster chains to infer topical similarities.

III. Methodology

Our overall framework (Figure 2) has the following main components: we first thread news articles into story chains, retrieve Twitter trends for every news report in a given story chain, detect and encode interaction patterns and finally use clustering and topic modeling to understand the topical differences among different interaction patterns. Each one of these components is described in detail in the following sections.

A. Story Chaining of News Articles

The chaining methodology is developed with the goal of identifying all documents related to a news story and to keep track of the news story as new documents arrive. Documents belonging to such a chain cover the same event and are ordered by time. Traditional clustering approaches can cluster together documents about similar events but are insufficient to separate out documents of each individual event. Thus we formalize an approach that chains together documents about an event as they appear, in order to build a narrative thread of that event. The algorithm operates in an incremental fashion wherein every new input article is analyzed as it arrives and is appended to already existing chain(s). This analysis involves a two-step process. In the first step, we compare an incoming article D_i to articles from the last n[1] days to identify the most similar articles and then designate candidate chains to which the current article can be attached to. If no similar articles are found, then a new chain is created with this article as a seed.

[1]Empirically, $n = 14$ (2 weeks) was found to be most effective.

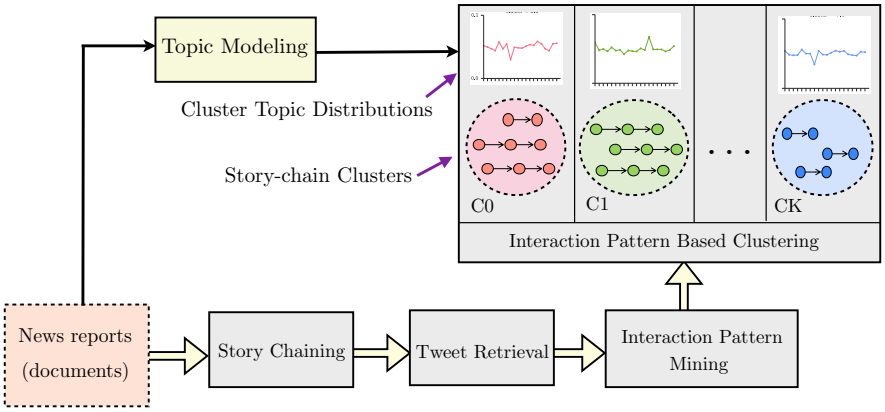


Fig. 2: Schematic of our interaction pattern mining framework.

Further, to assess if two documents are referring to the same underlying context, we calculate their *similarity scores* with respect to three features:

- textual features, denoted by $\mathcal{T}(D_i)$,

- spatial features, denoted by $\mathcal{L}(D_i)$, and

- actors, denoted by $\mathcal{A}(D_i)$.

The textual features are represented by the TF-IDF vector of the tokens present in the document. The spatial features are the set of locations mentioned within the text of an article. Every phrase/token identified as a location name by a named entity recognizer[2] (NER) is resolved into a <country, state, city> tuple with help of a geocoder based on probabilistic soft logic [21]. Details of the geocoding methodology are described in our prior work [22]. The weight l_{ij} of a location entity in the spatial feature vector represents the probability of appearance of location l_j in the document D_i.

$$l_{ij} = \frac{\text{Frequency of } l_j \text{ in } D_i}{\sum_k \text{Frequency of } l_k \text{ in } D_i}. \quad (1)$$

Similarly, the actors feature vector ($\mathcal{A}(D_i)$) represents the set of actors (persons, organizations as detected by NER) mentioned in an article. The weight a_{ij} of each element in the actors feature vector represents the probability of appearance of actor a_j in document D_i and is defined similar to (1).

The total weighted similarity measure between two documents, D_i and D_j, is then defined as follows

$$\text{sim}(D_i, D_j) \triangleq \alpha f(\mathcal{T}(D_i), \mathcal{T}(D_j)) + \beta f(\mathcal{L}(D_i), \mathcal{L}(D_j)) + \eta f(\mathcal{A}(D_i), \mathcal{A}(D_j)), \quad (2)$$

where f denotes a similarity metric such as cosine similarity or Jaccard's coefficient and the weighting coefficients α, β, η are chosen such that $\alpha + \beta + \eta = 1$. The textual similarity in this equation captures the similarity in terms of topical content whereas the spatial and actor vectors capture the similarity in terms of the event(s) described in the two documents. Thus the choice of the weights α, β, η control the relative coherence of two documents w.r.t. textual, spatial, and actor related features.

Eqn. 2 is used to obtain articles most similar to the current article from the past n days and thus a set of candidate chains to which the current article can attach could also be found. Once a candidate set of chains are found, in the second step, the candidate set is pruned based on the coherence of the article D_i with a story chain C_j. Here, coherence is calculated as the weighted sum of coherence between the spatial and actor feature vectors of an article and the spatial and actor feature vectors of a chain. The spatial feature vector $\mathcal{L}(C_j)$ and the actor feature vector $\mathcal{A}(C_j)$ of a chain are defined similar to (1) by considering all news articles in the chain as a single document. The coherence between a chain C_j and document D_i is defined as

$$\text{coh}(D_i, C_j) = \theta g(\mathcal{L}(D_i), \mathcal{L}(C_j)) + \phi g(\mathcal{A}(D_i), \mathcal{A}(C_j)),$$

where g is any similarity measure and the coefficients θ, ϕ are chosen such that $\theta + \phi = 1$. The spatial and actor feature vectors for a chain are then updated every iteration if there is any update i.e., any new document is added to the end of the chain.

The article D_i is attached at the end of all chains such that $\text{coh}(D_i, C_j) \geq \Gamma$, where Γ is the threshold which is used to tradeoff chain length and coherence. A higher value of Γ will cause the chains to be shorter but more coherent, and vice versa. If no chain passes similarity threshold Γ, then a new chain is created with this article. This two step process is repeated for every new article. Jason et al. present an evaluation of this chaining methodology in [23].

B. Retrieval of tweets related to news

Retrieving tweets related to a given news article is not trivial as tweets are comparatively very short (only 140 chars) and it is also necessary to find tweets associated with a news article *both before and after its publication* time in order to understand the information flow between news and Twitter. Sometimes, tweets mention the shortened URL of the actual news article thereby establishing an explicit connection indicating flow of information from traditional news media to the Twitterverse. However, such tweets are very few in number. On an average, from our experiments, we found that for a given news article, in our collection only about 5-6 tweets explicitly mention its URL (both shortened and unshortened forms were considered). On the other hand, certain news articles do also cite tweets or Twitter user names and hashtags in their content which can be used to find associated tweets appearing before the article got published. This again is only

[2]http://www.basistech.com/text-analytics/rosette/

a handful. Therefore, given the limitations of the API we used (Topsy), we resort to techniques of obtaining twitter count metrics by identifying tweets by keywords instead of techniques like topic filtering, BM25 and Rocchio methods. Specifically we follow a four-step process as illustrated below:

Step 1: Collect tweets mentioning a given URL. We harvest both the mentions of shortened and unshortened forms of the given URL.

Step 2: Extract top 10 keywords from the list of keywords obtained after tokenization and stopword removal of the text of tweets obtained in the earlier step. This list of keywords is combined with a set of entity words obtained by performing language enrichment on the news article (see Section III-A).

Step 3: Remove items from the previously obtained list of keywords plus those entities that are common to other articles in the same chain as the current article. This is necessary because all news articles in a story-chain share some common topics and so will their corresponding Twitter activity. Thus it is important that we extract information unique to a particular news report versus that of other news reports in a chain. This step is necessary as it helps us study news-Twitter interaction at an individual article levels without having to consider inter-dependencies.

Step 4: Download hourly count metrics for each element in the keywords plus in the entities list obtained in the last step. The Twitter count metric download is limited to the time window of $[t_0-7, t_0+7]$ days, where t_0 is the article publication date.

C. Identifying Interaction Patterns between News and Tweets

At this point, we have news articles grouped together to form story chains and for each news report in a story chain, we have its corresponding Twitter activity profile. In this section, we discuss how interaction is defined for a single news article and then use this information to define interaction patterns for a whole chain.

Peaks in Twitter activity showcase interestingness and can indicate either inflow of information from another source or possible triggers for outflow of information to a different source. We assume that the presence of peaks in the Twitter activity is a good milestone to use to posit interactions between news and Twitter. For all our experiments, we assume the interaction is only between news wires and tweets and that there is no other third source. The algorithm for peak detection [24] is detailed in Algorithm 1. Peaks are defined to be those points in time where the corresponding value is higher than its immediate surrounding (± 3 hours) and the difference is much higher than the standard deviation of the entire series. Peaks that appear close to the article publish time have higher possibility to influence or get influenced by the news article depending on whether they happens before or after the article is published. Hence, the *net influence is not only based on time lag between the news article and peak but also the actual peak value*. In short, we define the influence weight of a Twitter peak to be directly proportional to its peak value and inversely proportional to the time lag between the peak and the publication time of the news article. The influence weights of pre- and post- article publish time peaks are summed up separately to capture the net incoming influence $\mathcal{W}^{\text{pre}}$ and the net outgoing influence $\mathcal{W}^{\text{post}}$ of a news article as:

Algorithm 1 Peak Detection in Twitter activity profile

```
procedure DetectPeaks(y, threshold)
    m = std(y); y_m = mean(y)
    inds = [] as peak position array
    for y_i in y do
        if y_i satisfies the following constraints then
            (1) y_i > y_{i+1} and y_i > y_{i-1}
            (2) y_i > y_m
            (3) min(y_i - y_{i+1}, y_i - y_{i-1}) > m
            (4) i > inds.last() + threshold
            inds.append(i)
    return inds
```

$$\mathcal{W}^{\text{pre}} = \sum_{s \in S_{\text{pre}}} \frac{v_s}{t_A - t_s}, \quad \mathcal{W}^{\text{post}} = \sum_{s \in S_{\text{post}}} \frac{v_s}{t_s - t_A}, \tag{3}$$

where

– t_A is the time of publication of news article A.

– S_{pre} is the set of peaks detected before t_A.

– S_{post} is the set of peaks detected after t_A.

– t_s is the occurrence time of the peak s and v_s is the peak value after normalizing the Twitter activity profile so that values range from 0 to 1.

Next, using the net incoming and outgoing influence weights, we define four interaction states in which a news article can be in:

- N : Here, the direction of information flow is predominantly from News to Twitter. Thus, $\mathcal{W}^{\text{pre}}$ is not significant whereas $\mathcal{W}^{\text{post}}$ can be significantly higher compared to $\mathcal{W}^{\text{pre}}$.
- T : This state indicates Twitter is the major information source and the flow from Twitter to news is significant as compared to the reverse flow. Mathematically, $\mathcal{W}^{\text{pre}}$ is significant and $\mathcal{W}^{\text{post}}$ is not significantly higher than $\mathcal{W}^{\text{pre}}$.
- B : State B represents bi-directional information flow between news and Twitter. Here, both $\mathcal{W}^{\text{pre}}$ and $\mathcal{W}^{\text{post}}$ can be significant.
- E : This state denotes absence of any significant information flow, i.e., both $\mathcal{W}^{\text{pre}}$ and $\mathcal{W}^{\text{post}}$ are insignificant.

Formally, these set of states are defined in Equation. 4.

$$\text{State}(D_i) = \begin{cases} N, & \text{if } \mathcal{W}^{\text{pre}} < \rho,\ \mathcal{W}^{\text{post}} \geq (1+\lambda)\mathcal{W}^{\text{pre}} \\ E, & \text{if } \mathcal{W}^{\text{pre}} < \rho,\ \mathcal{W}^{\text{post}} < (1+\lambda)\mathcal{W}^{\text{pre}} \\ T, & \text{if } \mathcal{W}^{\text{pre}} \geq \rho,\ \mathcal{W}^{\text{post}} < (1+\lambda)\mathcal{W}^{\text{pre}} \\ B, & \text{if } \mathcal{W}^{\text{pre}} \geq \rho,\ \mathcal{W}^{\text{post}} \geq (1+\lambda)\mathcal{W}^{\text{pre}} \end{cases} \tag{4}$$

Here, ρ is the significance threshold for $\mathcal{W}^{\text{pre}}$ signifying the level of Twitter activity before the article publish time. λ is the significance threshold for the difference between $\mathcal{W}^{\text{pre}}$ and $\mathcal{W}^{\text{post}}$, which corresponds to the $\%$ increase in Twitter activity after article publish time. Fig. 3 shows the sectors represented by each of these four interaction types in a cartesian plane defined by $(\mathcal{W}^{\text{pre}}, \mathcal{W}^{\text{post}})$. This figure also shows *typical* Twitter

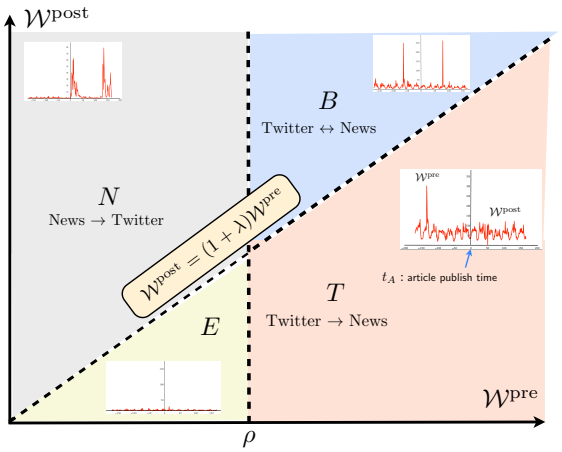


Fig. 3: Geometric interpretation of four interaction types N, T, E, B in a cartesian plane defined by $\mathcal{W}^{\text{pre}}$ and $\mathcal{W}^{\text{post}}$.

activity profile(s) around article publish time corresponding to the four interaction states.

Once every article in a chain is assigned a state based on Equation. 4, the interaction pattern of a chain can be defined as the concatenated string of interaction states of each individual article of the story chain in temporal order. For example, for the chain in Fig. 1 the interaction pattern is "NTNTBNNE". Each individual character in the string represents the interaction state of the corresponding news article with respect to its Twitter activity profile. This type of encoding is referred to as a *qualitative encoding*. Also every chain can be represented as a two-dimensional real valued vector where one dimension represents the $\mathcal{W}^{\text{pre}}$ values of each article in the chain and the other represents $\mathcal{W}^{\text{post}}$ of each article. This form of encoding will be referred to as a *quantitative encoding*.

D. Clustering of Interaction Patterns

In this section, we present two approaches for clustering the story chains using their interaction patterns. Clustering is performed using both qualitative and quantitative encoding of interaction patterns as both offer different advantages.

Clustering via qualitative encoding – Using the qualitative encoding, every chain is represented as a string of labels (e.g., "TTEBNEB"), each label corresponding to the interaction states of articles in the chain. We then use string edit distance metrics to calculate the difference in interaction patterns among two story chains. As story chains can differ widely in lengths, we collapse repetitive letters into one to reduce the encoding size differences among different chains so that the string edit distance metrics are more effective. Thus for the example in Fig. 1, the collapsed representation is "NTNTBNE". We explore distance based k-medoids clustering with a setting of 5 clusters and the following possible string edit distances:

• Levenshtein distance [25], which is the edit distance between two sequences and is defined as the minimum number of single character edits to change one sequence into the other.

• Jaro-Winkler distance [26], which is also a type of edit distance that was developed in the area of record linkage. It uses the idea that differences at the start of a string are more significant than edits at the end of a string.

• Ratcliff-Obershelp pattern recognition [27] which computes the similarity between two strings as the doubled number of matching characters divided by the total number of characters in the two strings.

We select the Jaro-Winkler distance for our analysis henceforth as it has a lower intra-cluster distance for $K = 5$ clusters.

Clustering via quantitative encoding – clustering based on the qualitative encoding suffers from certain disadvantages. The primary disadvantage is that the string edit distance metrics are quite sensitive to string length. Collapsing the string encodings as we descibed above reduces the effect of this problem a little but does not get rid of it in its entirety. For this reason, we encode chains as a two-dimensional vector of values of $\mathcal{W}^{\text{pre}}$ and $\mathcal{W}^{\text{post}}$ of each individual article in the chain. This form of encoding allows us to apply multi-dimensional dynamic time warping (DTW) to help compare the differences in interaction patterns of two chains of different lengths. DTW [28] finds the optimum alignment between two sequences of observations by warping the time dimension with certain constraints, thus allowing comparison of two sequences of different lengths.

E. Topic Modeling

Thus far now, we clustered story chains by employing the interaction patterns between news and tweets. To identify hidden topics underlying each cluster of story chains and explore if certain specific interaction patterns show interests in specific topics, we apply topic modeling algorithms on the news report collections. Specifically, we use latent Dirichlet allocation [29] to generate distributions over words for each topic (and also obtain the proportions of topics distributed in a document). Then we define the weights over the mixture of topics for one cluster by:

$$\mathbf{C}_{j,k} = \frac{\sum_{d_{ij} \in c_j} n_{d_{ij}} \theta(d_{ij}, k)}{\sum_{d_{ij}} n_{d_{ij}}}, \tag{5}$$

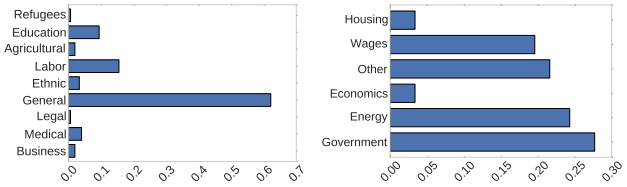
where n_{d_ij} refers to the frequency of d_i in cluster C_j and $\theta(d_{ij}, k)$ refers to the topic proportions for this document.

IV. Dataset and Experiments

We show the results of our framework on real data from Brazil during the period from Nov. 2012 to Sep. 2013. This period was chosen due to unusually high social media activity and news coverage around Brazilian mass protests (also known as the "Brazilian Spring"). We collected news reports in this period from top three leading news agencies in Brazil– *O Globo, Estadao,* and *Jornal do Brasil*. The news chaining algorithm proposed in Section III-A was applied to this corpus yielding 13,529 chains, out of which chains with a minimum length of 3 were considered resulting in 9933 chains. For every news report in these story chains, geo-targeted Twitter activity profiles (limited to Brazil) were collected using the Topsy api[3] as described in Section III-B.

In addition to this data, we also have access to a human curated list of civil unrest events that happened during this period. This list, called the Gold Standard Report (GSR) described in Ramakrishnan et al. [12], contains news reports for each event from the three major news sources. The GSR

[3]http://api.topsy.com/doc/



(a) population (b) event type

Fig. 4: Population and Event type Distribution in GSR chains.

TABLE I: Statistical properties of GSR and Non-GSR chains.

Category	Avg-Time-Lag(hour)	% of Twitter starts
GSR Chains	10.95	40%
Non-GSR Chains	5.26	73%

TABLE II: % of Twitter, News starts for GSR story-chains

Category	% News starts	% of Twitter starts
Housing related protests	100%	0%
Other (religious & cultural)	60%	40%
Govt. Policies	23%	77%
Medical	74%	26%
Agriculture	100%	0
General Population	30%	70%

also contains human-annotated information about the type of event and the group of people protesting as shown in Fig. 4. We separate the set of story chains into two categories: GSR chains (containing at least one GSR reported civil unrest event) and Non-GSR chains (with no GSR reported event). As GSR chains have more information, this segregation further highlights the differences in interaction patterns in terms of event type and population, which is not available otherwise for Non-GSR chains.

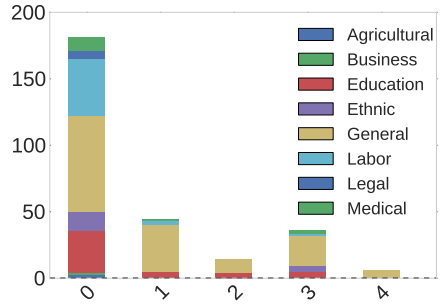
V. Results

We present four main results as follows:

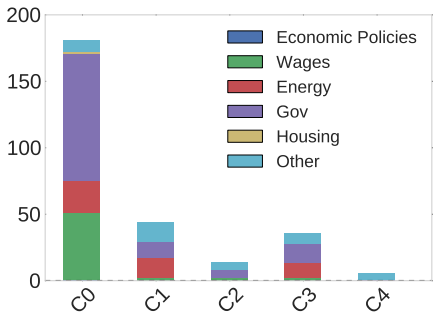
*1) **Comparison of GSR with Non-GSR chains**:* Statistical properties of GSR and Non-GSR chains are depicted in Table I. Here the column "% of Twitter Starts" refers to the percentage of chains where the interaction pattern starts with a information flow from Twitter to news i.e, chains with the qualitative encoding starting with "T" or "B". Avg-Time-Lag refers to the average time difference between between any two consecutive news reports in a story chain. Some interesting facts from this table are: (1) the average time lag between two consecutive news reports in story chains is less than half a day. This is mainly because the period for which our data spans includes events such as the major soccer tournaments of the Confederations cup and mass protests known as the Brazilian Spring[4]. (2) A lot of protest events, specifically approximately 40%, have Twitter starts indicating the presence of precursory signals in Twitter. Table II shows the breakup of the GSR chains in terms of event types and population which show that Govt. related, General Population and Other (Religious, Cultural, Social) events have a significant % of Twitter starts.

*2) **Difference in interaction pattern for GSR chains**:* We cluster the GSR chains into $K = 5$ clusters based on both qualitative and quantitative encoding(s) as described in Section III-D. As detailed in the previous section, each GSR chain can be attached to a set of event types and populations. Fig. 5 gives the distributions of event type and populations for each of the 5 clusters of GSR chains. Both the clustering using the Jaro-Winkler distance metric on the qualitative encoding and the DTW metric on the quantitative encoding yield similar results and thus only one of them is reported here.

[4]http://en.wikipedia.org/wiki/2013_protests_in_Brazil



(a) Population Distribution of Clusters (K-Medoids)



(b) Event Type Distribution of Clusters (K-Medoids)

Fig. 5: Stacked bar plots showing difference in event type and population distributions for different interaction pattern cluster. Cluster 4 patterns are mainly found in "other" related events and never in "energy". Similarly, "medical" population type can only be seen in events having interaction patterns from clusters 0 and 3. Also events with "labor" as population fall primarily in cluster 0.

The variability in different population distributions across clusters is clearly evident. For instance, Cluster 0 has a greater share of story chains with population(s) such as "education", "ethnic" and "labor". Within the encoded sequences of this cluster, the most common sub-sequence at the beginning of the interaction pattern is ("NBNBT") and ("E"). This indicates that at the beginning of such story chains, there is a news report coming out first and then Twitter discussion starts. Also, the alternative sub-patterns such as "NB" indicate that the discussion in Twitter involving these populations are not consistently active. In this cluster, chains encoded as "E" (with insignificant activity in Twitter) have population distribution over "agricultural", "legal" and "business". Clusters 2 and 3 are predominantly from "general population" with some share of "medical" and their encoded patterns look like "TBE", "TNT", "BNT" etc., which implies that for such clusters in general the Twitter is the first to break the story.

Regarding event types, there is diversity in terms of the distribution of "energy", "govt. related" and "wages" related protests. Cluster 0 has a higher percentage of "wages" and

TABLE III: Important words in topics inferred by LDA

Topics	Words
0. Economy	economia,brasil,milhao,bilhao,banco,mes,produto
1. Others	paulo,brasil,noticia,zap,portal,short,url,urllonga,anunciar
2. Government	governo,paulo,publico,projeto,presidente,dever,direito
3. Local Event	falar,saber,ceara,querer,mundo,nacional,clube,passar,fortaleza
4. Protest	manifestante,protesto,rio,feira,pessoa,avenida,policia
5. Entertainment	brasil,paulo,cultura,show,mostrar,cinema,gaga,filme
6. Internet	mail,cadastrar,login,senha,cpf,querer,abaixo,guardar
7. Business	empresa,podar,energia,mercado,setor,central,industria,servico
8. Crime	matar,pessoa,morrer,atingir,regiao,noticia,morte,violencia,vitima
9. Medical	rio,papa,janeiro,medico,indio,maracana,francisco,hospital
10. International	internacional,america,eleicao,brasil,unidos,pais
11. Judicial	partido,presidente,dilma,federal,paulo,ministro
12. Advertisement	publicidade,brasil,rio,jornal,tolipan,heloisa,cultura
13. Transportation	paulo,policia,onibus,policial,ataque,capital,veiculo
14. Sports	copa,futebol,jogo,esporte,brasileiro,selecao,paulo
15. Geography	rio,chuva,cidade,casa,regiao,pessoa,janeiro,
17. Police	policia,policial,militar,morte,matar,crime,caso,preso
18. Local Event	boate,maria,santo,incendio,pessoa,sul,tragedia,kiss
19. Technology	tecnologia,ciencia,sol,vinho,anna,ramalhar,ambiental,cultura

TABLE IV: Top topics for clusters

Cluster ID	Frequent Sub-patterns	Top Topics
C0	"NBNBTNTN", "NTNTN"	Local Events
C1	"NT", "NTNT"	Local Events
C2	"TNT"	Local Events, Ads, Technology
C3	"T", "TB"	Others, Protest, Sports
C4	"TNENT", "TEB"	Protest, Government, Entertainment

"government" related protests. This cluster, in terms of population, consists of "education", "ethnic", and "labor". Cluster 2 and 3 exhibit different event types. Cluster 3 has more story chains about "energy" and cluster 2 involves "gov" and "others". Both of these two clusters have general population related patterns and more starters with Twitter. Cluster 4 has more proportion in "other" where also some proportion of "general population" events are found.

*3) **Topic Variability in Interaction Patterns**:* In order to understand the generic differences in topics exhibited by different clusters, we applied topic modeling to the documents in our dataset and used them to calculate topic distributions for each cluster as described in Section III-E. For this set of experiments, we included both GSR and Non-GSR chains. Table III shows the results from LDA. The topic labels were assigned manually by our domain expert. Fig 6 (next page) gives a general description of distributions of 20 topics over 5 clusters. Referring to Table IV to get an idea of what each topic talks about, we can see that the distribution differences across the clusters in Fig 6 can be intuitively explained.

The discussions related to local events (natural and man-made) peaked in news before Twitter, though there could have been a few tweets about the event earlier than news. This can be seen by the appearance of N in the frequent sub-patterns in such story chains. This inference is probably local to the type of events (fire accident) that happened during the period of analysis and of the geographical region (Brazil) studied in our dataset. Moreover, we see reciprocity between news and Twitter via subsequent alternating appearances of B, N and T states. In contrast, stories related to sports, protests and advertisements tend to become trends earlier on Twitter followed by back and forth interactions with the traditional news media (as seen by the frequent sub-patterns for the chains in clusters 2, 3 and 4).

*4) **Which one is the main influencer?**:* We define the influence weight of a story chain as the average of the difference of pre- and post- influence weights, $\frac{\sum_i (\mathcal{W}_i^{\text{pre}} - \mathcal{W}_i^{\text{post}})}{n}$, where the summation is over n, the number of articles in a chain. This influence weight is used to identify the main influencer for a story chain i.e., which direction the information flow is predominant over its course. If the influence weight is positive, larger absolute value implies Tweets are more active in this story chain and the flow of information is mostly from Twitter to news. If the weight is negative, the larger absolute value indicates news is mostly earlier than Tweets in reporting the sub-events within a story chain. Influence weight close to zero indicates significant reciprocal interaction between news and Twitter over the course of the story chain.

Table V (shown on next page) lists the top most chains with the highest positive, and negative influence weights as well as chains with approximately zero weights. We also present the corresponding interaction patterns, and a brief description of the story chain. For instance, story chain SC1, whose main influencer is Twitter talks about a student organized protest at the door of a church against the appointment of pastor Marco Feliciano at the presidency of Human rights of the federal chamber in Sao Paulo, Brazil. Story SC5 talks about the arrival of tennis player Rafael Nadal for Brazil's only ATP tournament after 8 long years causing much hype in both traditional and social media. Overall, we can see that news domains track events related to politics and local events, earlier than Twitter, whereas Twitter is quicker in capturing information about social events and famous figures.

VI. Conclusions

In this paper, we presented a new framework for discovering the direction of information flow over time across two heterogeneous information media - news and Twitter. This lets us uncover the interaction patterns overs stories consisting of chronologically chained news articles. We tested the proposed interaction pattern framework on real data from Brazil and we found that both Twitter and traditional News media have variable influences on different topics. Twitter as a social network platform serves as a fast way to draw attention from public for many social events such as sports news whereas news media is quicker to report events regarding political, economical and business issues. Certain topics were found to have similar influence from both Twitter and News media.

For future work, we would like to not only use the temporal trends of Twitter but also the textual features of each individual tweet related to a news report. Also another interesting direction is to distinguish between explicit vs implicit interactions between news and Twitter.

Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

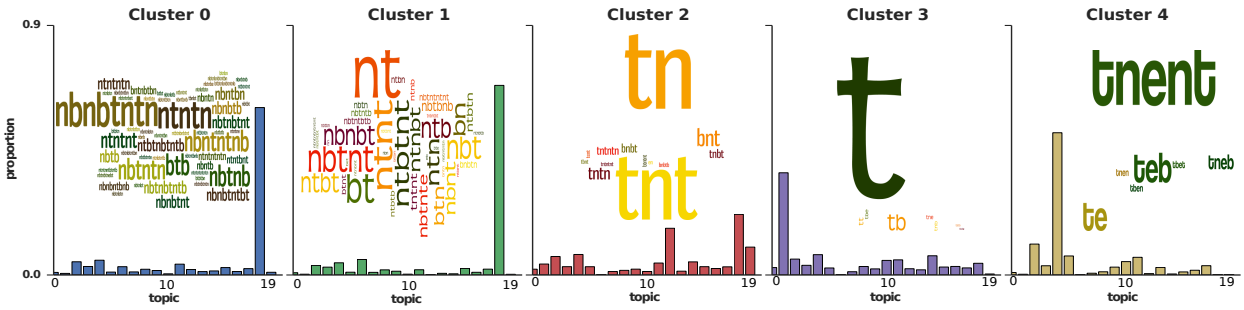


Fig. 6: Topic distributions of each interaction pattern cluster. The X-axis labels refer to topic number as given in Table III. Also for each cluster a pattern cloud of the most significant interaction patterns found in the cluster is shown. For example, Cluster 3 has high proportion of topic 1 and mainly has story chains where information flow is mainly from Twitter to news.

TABLE V: Story Chains with Interaction Patterns and Main Influencer

StoryChain-ID (SC)	Interaction Pattern	Influence Weight	Main Influencer	Story Summary
SC1	TT	0.514	Twitter	"Marco Feliciano enfrenta protesto na porta de igreja"
SC2	TN	0.48	Twitter	"25%Teachers are on strike. Government denies."
SC3	NNNNBNTBN	-0.422	News	"Fire in Kiss Nightclub in Santa Maria "
SC4	NBNNTN	-0.405	News	"Governor Genro decress official mourning"
SC5	TTTNN	5.0e-05	Both	"After 8 years, Nadal back to Brazil with high investment and large team"
SC6	NNTNTTNTN	-1.7e-04	Both	"Nissan sells more than 100 thousand first"

REFERENCES

[1] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proc. of the 19th WWW*, 2010.

[2] D. Kumar, N. Ramakrishnan, R. F. Helm, and M. Potts, "Algorithms for storytelling," *IEEE Transactions on Knowledge and Data Engineering*, 2008.

[3] M. S. Hossain, P. Butler, A. P. Boedihardjo, and N. Ramakrishnan, "Storytelling in entity networks to support intelligence analysts," in *Proc. of the 18th ACM SIGKDD*, 2012.

[4] C. Faloutsos, K. S. McCurley, and A. Tomkins, "Fast discovery of connection subgraphs," in *Proc. of the 10th ACM SIGKDD*, 2004.

[5] M. S. Hossain, M. Akbar, and N. F. Polys, "Narratives in the network: interactive methods for mining cell signaling networks," *Journal of Computational Biology*, 2012.

[6] M. S. Hossain, J. Gresock, Y. Edmonds, R. Helm, M. Potts, and N. Ramakrishnan, "Connecting the dots between pubmed abstracts," *PloS one*, 2012.

[7] D. Shahaf and C. Guestrin, "Connecting the dots between news articles," in *Proc. of the 16th ACM SIGKDD*, 2010.

[8] D. Shahaf, C. Guestrin, and E. Horvitz, "Trains of thought: Generating information maps," in *Proc. of the 21st WWW*, 2012.

[9] D. Shahaf, J. Yang, C. Suen, J. Jacobs, H. Wang, and J. Leskovec, "Information cartography: creating zoomable, large-scale maps of information," in *Proc. of the 19th ACM SIGKDD*, 2013.

[10] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proc. of the 15th ACM SIGKDD*, KDD '09, (New York, NY, USA), pp. 497–506, ACM, 2009.

[11] F. Chierichetti, J. Kleinberg, R. Kumar, M. Mahdian, and S. Pandey, "Event detection via communication pattern analysis," in *8th International AAAI Conference on Weblogs and Social Media*, 2014.

[12] N. Ramakrishnan, P. Butler, S. Muthiah, *et al.*, "'Beating the news' with EMBERS: Forecasting civil unrest using open source indicators," in *Proc. of the 20th ACM SIGKDD*, 2014.

[13] Y. Hu, A. John, F. Wang, and S. Kambhampati, "Et-lda: Joint topic modeling for aligning events and their twitter feedback," in *Proc. of the 26th AAAI Conference on Artificial Intelligence*, 2012.

[14] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang, "Transferring topical knowledge from auxiliary long texts for short text clustering," in *Proc. of the 20th ACM CIKM*, 2011.

[15] J. Hopcroft, T. Lou, and J. Tang, "Who will follow you back?: Reciprocal relationship prediction," in *Proc. of the 20th ACM CIKM*, CIKM '11, (New York, NY, USA), pp. 1137–1146, ACM, 2011.

[16] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: Understanding microblogging usage and communities," in *Proc. of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, 2007.

[17] S. Petrovic, M. Osborne, R. McCreadie, C. Macdonald, I. Ounis, and L. Shrimpton, "Can twitter replace newswire for breaking news?," in *Proc. of the 7th ICWSM*, ICWSM, 2013.

[18] I. Subasic and B. Berendt, "Peddling or creating? investigating the role of twitter in news reporting," in *Advances in Information Retrieval*, 2011.

[19] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?," in *Proc. of the 19th WWW*, 2010.

[20] J. Sankaranarayanan, H. Samet, *et al.*, "Twitterstand: news in tweets," in *Proc. of the 17th international conf. on advances in geographic information systems*, 2009.

[21] A. Kimmig, S. Bach, M. Broecheler, B. Huang, and L. Getoor, "A short introduction to probabilistic soft logic," in *Proc. of the NIPS Workshop on Probabilistic Programming: Foundations and Applications*, 2012.

[22] S. Muthiah, B. Huang, J. Arredondo, D. Mares, *et al.*, "Planned protest modeling in news and social media," in *Proc. of the 27th IAAI*, 2015.

[23] J. Schlachter, A. Ruvinskya, L. Asencios Reynoso, S. Muthiah, and N. Ramakrishnan, "Leveraging topic models to develop metrics for evaluating the quality of narrative threads extracted from news stories," in *Proc. of the 6th International Conference on Applied Human Factors and Ergonomics*, AHFE, Elsevier, 2015.

[24] M. Duarte, "Notes on scientific computing for biomechanics and motor control," 2015.

[25] V. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, 1966.

[26] M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida," *Journal of the American Statistical Association*, 1989.

[27] P. E. Black, *Ratcliff /Obershelp pattern recognition*. 2004.

[28] T. Vintsyuk, "Speech discrimination by dynamic programming," *Cybernetics*, 1968.

[29] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, 2003.