International Journal of Disaster Risk Reduction 38 (2019) 101204

Contents lists available at ScienceDirect

International Journal of Disaster Risk Reduction

journal homepage: www.elsevier.com/locate/ijdrr

Exploring the emergence of influential users on social media during natural disasters

Yang Yang[a,1], Cheng Zhang[b,*,1], Chao Fan[b], Wenlin Yao[a], Ruihong Huang[a], Ali Mostafavi[b]

[a] *Department of Computer Science & Engineering, Texas A&M University, USA*
[b] *Zachry Department of Civil Engineering, Texas A&M University, USA*

ARTICLE INFO

Keywords:
Social media
Public information and warning
Twitter
Influential user
Hurricane Harvey
Disaster

ABSTRACT

Public information and warning represent key components of society's disaster preparedness, as they complement planning and operational coordination. Social media has increasingly become a means by which users spread useful public safety and emergency information during these events. The emergence of influential users plays a critical role in the improvement of online social networks for widespread information diffusion. This study focuses on the emergence of influential individual users on Twitter and explores the reason why they achieved a significant number of followers during one of the most devastating disasters: Hurricane Harvey. The diverse patterns of increases in followers for the emerging influencers may be separated into several types. Among those types, Twitter users who posted objective disaster-related information in clear language and a consistent fashion tended to show an increase in followers and emerged as influential users. Additionally, professionals (from the media industry, public agencies and departments, and research organizations) made significant contributions to public information and warning during the hurricane. The findings highlight the common attributes of emerging influential social media users and the crucial role they play in the self-organized dissemination of disaster information.

1. Introduction

Natural disasters such as earthquakes and hurricanes often prove devastating for the people who live where they occur. Access to timely and reliable information about the situation, protective and preventative measures, and critical lifesaving and life-sustaining efforts can decrease mortality [1]. Therefore, public information and warning represent key components of society's disaster preparedness, as they complement planning and operational coordination [1,2]. Social media comprises a part of the societal infrastructure and increasingly has been used to improve upon public information and warning during disasters. Social media enables multiple paths of communication from a single source. Therefore, it contributes to the efficiency of information dissemination [3–5]. Information made available on online social networks—similar to other complex networks—occurs through different diffusion processes [6,7]. An important diffusion process is information propagation through influential individuals (also called ‘influencers’ or ‘hubs,’ [8,9]), who are users (i.e., nodes in the network) with a large number of followers (links). In the context of a disaster or an emergency event, certain users quickly emerge by gaining many followers; they become information hubs for the communication of reliable information to the public. Emerging influential users play an important role in information dissemination during disasters. For example, a meteorologist in the Harris County Flood Control District (HCFCD) who actively posted flood-related information and answered questions during Hurricane Harvey increased his number of Twitter followers from 2633 to 13613 during seven days in August 2017. To recognize the meteorologist's contribution, the mayor of Houston designated May 2nd as a day of honor in his name [10]. Despite the important role they play during disasters, the existing literature has little information on the characteristics of emerging influential social media users.

The lack of research on the emergence of influential users contributes to a gap in understandings about online social networks. Who are the users that have a significant influence on the dynamics of online social networks during extreme events? The online social network structure is highly dynamic regarding link creation and deletion [11,12], and frequently is referred to as “link prediction” in the literature [13]. Link prediction studies primarily have focused on

[*] Corresponding author.
E-mail addresses: yangyangsandy@tamu.edu (Y. Yang), czhang@tamu.edu (C. Zhang), chfan@tamu.edu (C. Fan), wenlinyao@tamu.edu (W. Yao), huangrh@cse.tamu.edu (R. Huang), mostafavi@tamu.edu (A. Mostafavi).
[1] Yang Yang and Cheng Zhang have equal contribution to this paper.

https://doi.org/10.1016/j.ijdrr.2019.101204
Received 24 January 2019; Received in revised form 1 June 2019; Accepted 2 June 2019
Available online 08 June 2019

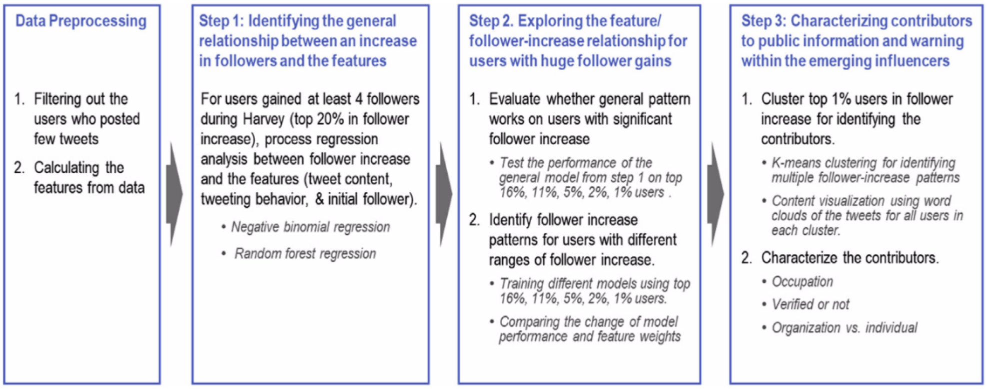
Data Preprocessing
1. Filtering out the users who posted few tweets
2. Calculating the features from data
Step 1: Identifying the general relationship between an increase in followers and the features
For users gained at least 4 followers during Harvey (top 20% in follower increase), process regression analysis between follower increase and the features (tweet content, tweeting behavior, & initial follower).
• Negative binomial regression
• Random forest regression
Step 2. Exploring the feature/ follower-increase relationship for users with huge follower gains
1. Evaluate whether general pattern works on users with significant follower increase
• Test the performance of the general model from step 1 on top 16%, 11%, 5%, 2%, 1% users .
2. Identify follower increase patterns for users with different ranges of follower increase.
• Training different models using top 16%, 11%, 5%, 2%, 1% users.
• Comparing the change of model performance and feature weights
Step 3: Characterizing contributors to public information and warning within the emerging influencers
1. Cluster top 1% users in follower increase for identifying the contributors.
• K-means clustering for identifying multiple follower-increase patterns
• Content visualization using word clouds of the tweets for all users in each cluster.
2. Characterize the contributors.
• Occupation
• Verified or not
• Organization vs. individual

Fig. 1. Overview of the methodology.

modeling the dynamics of direct links (e.g., following and friending, [14,15]) and the indirect links that result from information flow (e.g., retweeting and replying, [16,17]) between nodes, as well as the interactions between direct and indirect links (e.g., the tweet-retweet-following process, [8,18,19]). These studies have produced a variety of techniques that help to predict the generation and destruction of links that contribute to the evolution of networks (e.g., recommendations of Twitter accounts that people may want to follow [20,21]). However, they have provided limited insights into the human behaviors that inform the dynamics of online social networks.

Link prediction studies typically have been based on the assumption that mutually similar nodes will link to the same node or other mutually similar nodes. That assumption helps to predict the unknown or future links based on known links but fails to uncover the less obvious reasons for link creations. Consequently, existing studies have focused heavily on ways in which to measure the similarity between nodes regarding the user profile, the content of social media posts, and network topology [13,21]. However, very little research has examined the underlying reasons for which a group of nodes would link to one another [20,22]. Moreover, no studies have been conducted that specifically focus on emerging influential users (information hubs) or identify the behaviors that trigger the public's recognition and promotion of them during extreme events.

Link prediction techniques to date could result in the recommendation of meteorologist above to a Twitter user during Hurricane Harvey but would not be able to account for his popularity. During uneventful times, people could have a variety of reasons for the creation of links on social networks (e.g., family and friends, shared hobbies or topics of mutual interest). However, during extreme or emergencies, the emergence of influencers suggests that the public has shared understandings that would contribute to the promotion of certain users. The public's need to find and share reliable information related to disaster situations could prove to be an invaluable resource that could be deployed to improve public information and warning.

This study addresses the gap in the literature with a focus on influential Twitter users during Hurricane Harvey. The reason why the authors focused on Hurricane Harvey is threefold. First, Harvey is one of the costliest hurricanes in U.S. history [23], which showed its huge impact. More importantly, Hurricane Harvey is referred to as “the U.S.'s First Social Media Storm” [24] because people proactively used social media to process disaster response activities, including disseminating situational information and seek rescues; many influential users emerged during this disaster (e.g., the meteorologist in the HCFCD). The study examined the reasons why they gained the following that they did during the disaster. The following questions guided the research: 1) what are the features about tweet content and tweeting behavior that contribute to a significant increase in followers during disasters and 2) what are the social characteristics of the emerging influential users who contributed to disaster information dissemination?

The organization of the rest of this paper is as follows. Section 2 overviews the methodology. Section 3 shows the results and findings. Section 4 discusses the findings and shows the practical and theoretical value of this study. Section 5 concludes the paper.

2. Methodology

The overall goal of this study is to identify the emergence of influential individual users on Twitter and explore the reason why they achieved a significant number of followers during one of the most devastating disasters: Hurricane Harvey. The study began with a baseline comprised of an analysis of the relationship between a user's increase in the number of followers and the content of their tweets (step 1). As a means by which to distinguish emerging influencers from ordinary Twitter users, the study explored how the relative importance of different features varied for users with a greater increase in followers compared to the baseline user group (step 2). Finally, clustering analysis of the most influential users based on the features that corresponded to a significant increase in followers identified in the previous steps was used to identify and characterize the individuals who contributed to disaster information dissemination (step 3). Fig. 1 overviews the methodology.

2.1. Data collection and pre-processing

Data used in this research was collected from the time that Hurricane Harvey made landfall in Houston on August 25, 2017, to August 31, 2017. We collected tweets from all users whose location in their user profile was designated as Houston through Twitter's PowerTrack API. To best identify emerging influencers, established users with a large number of followers were excluded. We established a threshold (follower increase $>$ 89,196) that helped to distinguish users with a large following from others by using the method proposed by Ref. [25], calculating the sum of the average of initial followers and three times the standard deviation of initial followers. A total of 5662137 tweets posted by 117380 unique users was collected. In this dataset, each tweet posted by a specific user (including retweets and replies) included the real-time number of followers for each user. We constructed a timeline of follower changes and acquired the follower change number for each Houston-based user from this data.

To balance between the vast majority of users with lower follower increase and those who gain a substantial number of followers, we ranked the remaining 117380 users by the change in the number of followers during the studied period. As shown in Fig. 2, only 20% of users have a follower increase of 4 or above. To avoid the abundance of users with a neglectable follower increase, we decided to make the top 20% as our cutoff when analyzing the patterns of follower increase. Next, we filtered out any Twitter user who posted less than 55 tweets

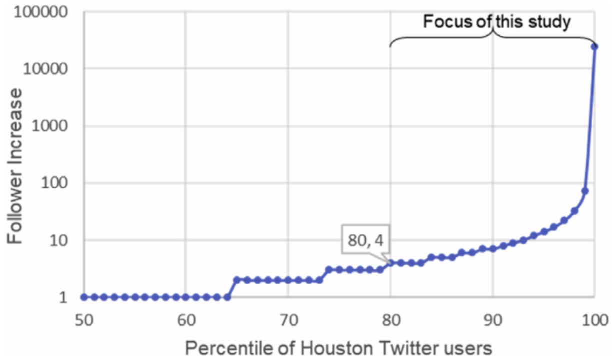

Fig. 2. Percentile of users ranked by follower increase (50%–100%).

Table 1
Distribution of top 20% users before and after filtering.

Percentile of Houstonian users by follower increase	Number of follower increase	Number of users	After filtering out users posted less than 55 tweets
20% and above	[4,+ ∞]	24127	11119
16% and above	[5, +∞]	19443	9633
11% and above	[7, +∞]	13439	7181
5% and above	[14, +∞]	5996	3486
2% and above	[33, +∞]	2402	1393
1% and above	[73, +∞]	1178	692

during Harvey. Most features of Twitter users in this study focused on Twitter content. Therefore, the features would not have proven accurate for the users who posted too few tweets. We established a threshold of 55 tweets by comparing the R^2 of the trained a series of random forest regression models during step 1 (to be introduced in detail in Section 2.3) with users who posted more than 40–100 tweets; performance of the random forest model peaked at 55 tweets. Performance of the models trained with other thresholds of tweet number was lower, due to either statistical noise or overfitting. Table 1 shows the distribution of top 20% users over different ranges before and after filtering.

2.2. Features that hypothetically influenced the increase in followers

The examined features that hypothetically influenced the increase in followers (hereafter referred to as "features") among Twitter users fell into the following three categories. The first category referred to content-related features.

- *Hurricane Harvey-related rate (hereafter referred to as "Related rate")* is the percentage of tweets that are related to Hurricane Harvey. We calculate this feature as the percentage of tweets that contained at least one of the event ontology keywords describing Hurricane Harvey. The event ontology keywords cover the following ten topics about Harvey: natural environment, preventative measure, help and rescue, utility and supply, business/work/school, utilities and supplies, fundraising and donation, flood control infrastructures, damage, transportation. To return all keywords into their root form for better matching results, they are processed by NLTK's implementation of Snowball stemmer before being used to match with the tweets [26].
- *Topic focus* indicated the similarities found among a given user's tweets. We converted each tweet of a user to term frequency-inverse document frequency (TF-IDF, [27]) vectors and calculated their pairwise cosine-similarity scores. Finally, we used the averaged pairwise similarity score for the topic focus feature [22].
- *Concreteness vs. abstract (hereafter referred to as "Concreteness")* measured how accessible the tweet was to readers, with one being the most abstract and five being the most concrete. A concreteness rating of 40 thousand common English words was used to measure the tweets' accessibility [28]. A concreteness rating for each tweet was obtained through the calculation of an average (from 1 most abstract to 5 most concrete) of words in the tweet. Words that did not appear in the ranking were given a score of 3. We calculated the concreteness rating for each user by taking the average of their tweets.
- *Informer rate* indicated how often a user disseminated non-original information, which is an indicator of how often outside information was used in a user's tweets [22]. The informer rate was the percentage of tweets by a user that contained URL, retweet, modified tweet, or head tip.
- *Meformer rate* indicated how often a user posted tweets about himself/herself, which is the percentage of tweets that contained at least one self-referential pronouns (e.g., I/me/my/we/our).
- *Originality rate* was the ratio of a user's original tweets (not retweets or replies) to total tweets. This feature is not the exact opposite of informer rate since the informer rate also included URLs, modified tweets, and head tips.
- *Sentiment score* measured the average positive or negative sentiment of a user's tweets. The average sentiment score for all tweets by a user was calculated using NLTK's implementation of VADER sentiment analysis [29].

The second category contained behavioral features and captured how consistently a user posted.

- *The average number of tweets posted per day* (hereafter referred to as "Tweets per day") was the average number of tweets posted by a user. An average number of tweets (including original tweet, retweet, and replies) a user posted during the seven days of study. The average number of tweets posted represented the primary behavior feature examined because it measured the amount of content the user created each day.
- *Number of replies* was the total number of replying tweets the user made, indicating the amount of interaction in which the user engaged with the public. The number of replies to other tweets the user made. This feature measured how often the user interacted with other users and represented an important means by which to differentiate accounts maintained by humans rather than bots.

The third category only contained one feature, which is a record of the *initial followers* each user had on August 25th, 2017, at the beginning of the analysis period.

Before doing further analyses, we first tested the features for multicollinearity, which may result in erroneous feature coefficients in regression models, using Variance Inflation Factors (VIF). As shown in Table 2, VIF for all features is less than 5, indicated by the rule of thumb for VIF to be of low multicollinearity. Therefore, we keep all the listed features in the following analysis.

Table 2
Variance inflation factor (VIF) of all features.

Feature	Variance Inflation Factor
Initial Follower	1.068818
Related Rate	1.193133
Topic Focus	1.358592
Originality Rate	1.377670
Number of Replies	1.644633
Informer Rate	1.184293
Meformer Rate	1.463350
Tweets per Day	1.521717
Sentiment Score	1.253889
Concreteness	1.358079

Table 3
Grid search parameters during optimization of the random forest model.

Hyperparameter	Range	Description
n_estimators	5,10, …,115,120	Number of decision trees
max_depth	1,2,3, …,19,20	Maximum depth for each decision tree
max_features	Square Root, Logarithmic, Auto (All Features)	Number of features to consider when looking for the best split

2.3. Step 1: identifying the general relationship between an increase in followers and the features of tweet content, posting behaviors, and initial follower

Step 1 of the study focused on the identification of the general relationships among an increase in followers, Twitter content, and posting behaviors. To be able to examine users with different levels of increases in followers, step 1 focused on users from Houston who gained at least four followers and ranked in the top 20% of increased followers during Hurricane Harvey. To better understand the contribution of these features and predict the follower increase, two regression models were employed to train the data and explore the importance of the features. Unless otherwise specified, the models used to describe the relationship between the increase in followers and the features of tweet content and tweeting behavior hereafter are referred to as 'models'.

The first model is generalized linear regression, which is a widely-used method to identify the variable importance. Specifically, because the dependent variable (i.e., follower increase) is an over-dispersed count variable (mean = 45.48, variance = 187,821), this study used the negative binomial regression [30]. Let Y_i be the dependent variable, which is the follower increase of each user i (i = 0,1,2,…, n) in the dataset. The probability mass function of Y_i is as follows:

$$\Pr(Y_i = y_i|\boldsymbol{x_i}) = \frac{\Gamma(y_i + \alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\frac{\alpha\mu_i(\boldsymbol{x_i})}{1 + \alpha\mu_i(\boldsymbol{x_i})}\right)^{y_i}\left(\frac{1}{1 + \alpha\mu_i(\boldsymbol{x_i})}\right)^{\alpha^{-1}},$$

$$y_i = 0,1,\ \ldots,$$

$$\ln(\mu_i(\boldsymbol{x_i})) = \boldsymbol{x_i}^T\boldsymbol{\beta} + \varepsilon$$

Where ε is the error term, which is a gamma-distributed random variable with mean one and variance α; $\Gamma(\bullet)$ is the gamma function. The predictor variables $\boldsymbol{x_i} = x_{1i},\ x_{2i},\ \ldots,\ x_{10i}$ are the normalized feature values introduced in Section 2.2. The standardized coefficients $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_{10}]$ are to be estimated. The variance and mean of Y are as follows, which shows the over-dispersed nature of the follower increase:

$$E(Y_i|\boldsymbol{x_i}) = \mu_i(\boldsymbol{x_i}),\quad Var(Y_i|\boldsymbol{x_i}) = \mu_i(\boldsymbol{x_i}) + \alpha\mu_i(\boldsymbol{x_i})^2$$

Therefore, α is also called the dispersion parameter. To estimate the parameters α and $\boldsymbol{\beta}$ using maximum likelihood estimation, we can write the likelihood function:

$$L(\alpha,\ \boldsymbol{\beta}) = \prod_{i=1}^{n}\frac{\Gamma(y_i + \alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\frac{\alpha\mu_i(\boldsymbol{x_i})}{1 + \alpha\mu_i(\boldsymbol{x_i})}\right)^{y_i}\left(\frac{1}{1 + \alpha\mu_i(\boldsymbol{x_i})}\right)^{\alpha^{-1}}$$

To calculate the parameters, we selected glm. nb from R, used logarithmic as the link function, and applied a limit of 1000 iterations to resolve the non-convergence for some features. We tested whether each feature correlated with the dependent variable and used the standardized coefficient $\boldsymbol{\beta}$ to measure how much weight each feature individually contributed to predicting the target.

In addition to the standardized coefficient obtained from the negative binomial model, we also used a random forest regression model, a non-linear ensemble model to conduct the training experiments and determine the relative weight of each feature in the prediction. The random forest regression model is also popular in identifying variable importance because it can handle a large number of variables as well as achieving high precision and fast computational time [31,32]. Because the range of our dependent variable extended across several magnitudes (10^0 to 10^4), we converted our target, follower increase, to log scale. For random forest, the score we used to indicate how much a feature contributes to the prediction is the mean decrease impurity (MDI) [33,44]. MDI is the average over all trees of the weighted sum of impurity decreases for all nodes in each tree where the feature is used [33]. With the training set of the top 20% of users in follower increase cross-validated into ten splits, we obtained the final feature weights by taking the average of the feature weights of the models trained with each split. We also optimized the hyperparameters of each random forest regression model with GridSearchCV function in the Scikit-learn package [34]. Table 3 shows the parameters used in the optimization process.

2.4. Step 2. exploring the feature/follower-increase relationship for users with huge follower gains

For the emerging influencers who gained a significant number of followers during Harvey, the factors that lead to follower increase might be different from that of the general public. Therefore, Step 2 of the analysis focused on exploring the relative importance of different features for Twitter users with significant follower gains. The hypothesis that informed step 2 was that the feature importance for follower increases of emerging influencers would be different from those in the model presented in step 1. Step 2 consisted of two sub-steps. The first sub-step tested whether the general model in step 1 could fit the users with the greatest increases in followers. We calculate the average of R-squared value and MSE of ten cross-validation models for users ranked top 20%, 16%, 11%, 5%, 2%, and 1% in follower increases, respectively, based on the predicted value of follower increase using the random forest models in step 1. The second sub-step fitted two regression models, negative binomial and random forest, to explore the relationship between follower increases and the features for users with different level of follower increases. We repeated the same optimization process as step 1 for each of the two sub-steps. The emerging influential users during Harvey are those who have a significant follower increase and have different patterns in the features related to follower increase compared to the general pattern in step 1.

2.5. Step 3: characterizing the contributors to public information and warning within the emerging influential users

To identify the contributors to public information and warning during Harvey within the users ranked top 1% in follower increase, who are typical examples of emerging influential users, we used the K-Means clustering algorithm to categorize the emerging influential users. K-Means clustering algorithm requires the number of clusters as input, and we used the elbow method [35] to determine an optimum number of clusters for the K-Means clustering algorithm. Elbow method determines the appropriate number of clusters within the dataset by looking at the curve of performance improvements as the number of clusters increases. For the performance metric of the clustering models, we used distortion, the sum of squared distances of points to the center of the cluster in each cluster [36]. By finding the "elbow" in the distortion curve, where a change of performance starts to decline, we can determine a point where increasing the number of clusters would not drastically increase the amount of variance any more, which is an indicator of the appropriate number of clusters within the dataset. Based on the clustering result, we determined different patterns of follower increase among the emerging influencers. The difference in the follower-increase patterns and tweet contents helps distinguish the contributors to public information and warning from other emerging influencers.

To further investigate the characteristics of the contributors, we analyzed their occupations through a manual annotation of their profiles. For each contributor, we first looked for direct clues indicating his/her/their occupation in the user profile, such as the name of the affiliation or description of the occupation. We then annotated the occupations of each contributor with seven labels: media, charity, community, medical, public sector, sports, and academia. We also annotated whether these accounts belong to an individual or an organization by looking at the account name and profile description. Finally, we collected the data on whether these accounts are verified by Twitter or not.

3. Results

This section introduces the findings corresponding with the three steps of analysis introduced in Section 2.

3.1. Step 1: identifying the general relationship between an increase in followers and the features

As is introduced in Section 2.3, two models (i.e., negative binomial and random forest regress model) are used to identify the relationship between follower increase and the features of tweet content, posting behaviors, and initial follower for the users ranked top 20% in follower increase. In the negative binomial model, all features have strong correlations with follower increase. Fig. 3a illustrates the ranking of the features according to the standardized coefficient and identifies the feature of the initial follower (the number of followers at the beginning of the analysis period) as the first outperforming the second feature (originality rate) by 101.7%. Among content features, originality rate and related rate proved to be the second and forth most important, while other content features such as informer rate, topic focus, and concreteness were ranked at the bottom. For behavioral features, 'tweet per day' ranked third, evidencing that the quantity of information posted also was a significant predictor of follower increase.

For the random forest regression model, Fig. 3b shows the relative feature weights of the trained model. A comparison of the ranking of the standardized coefficients (beta coefficients) from the negative binomial model and the ranking of the relative feature importance values in the random forest model served to determine the importance of each feature. Both models ranked by magnitude of the weights revealed that (1) initial follower has a significant influence on follower increase; and (2) related rate, originality rate, and tweets per day are important for a Twitter user to gain followers during Harvey. The performance of the proposed model is discussed as follows. The R squared score (R^2) of the random forest regression model was 0.383, and the mean squared error (MSE) was 0.129, both higher than the standard substantive significance level (0.26) used in the social sciences [37].

The results from the two regression models provided an understanding of the increase in Twitter followers during disasters. The number of initial followers ranked first by a large margin in both models. Related rate and originality rate were significant content features; the posting of original tweets (i.e., offering first-hand information about the disaster situation) was significant with regard to the increase in a given user's followers. One behavior feature, the average number of tweets per day ranked at the top, suggesting that posting a large amount of information represented an important means by which to gain followers. The 'meformer' proved negatively correlated with follower increase in the negative binomial model and was of relatively low significance in the random forest model. Content originality and relevance to the disaster, followed by the frequency of tweets, were the most important features for users to gain followers during Hurricane Harvey.

3.2. Step 2: exploring the feature/follower-increase relationship for users with huge follower gains

To investigate whether the general pattern of follower increase could be applied to the users with significant follower increase, we tested the general random forest model on users ranked the top 20%, 16%, 11%, 5%, 2%, and 1% in follower increase. As shown in Fig. 4, the performance of the general model declined significantly; the importance of the features that impacted follower increase for the emerging influential users was different from those for the ordinary users.

We proceeded to investigate the feature importance for Twitter users with different rankings with regard to follower increase (i.e. top 20%, 16%, 11%, 5%, 2%, and 1%) to identify the differences between the emerging influencers and ordinary users based on negative binomial regression and random forest regression. First, the standardized coefficients of the negative binomial were calculated. The results showed that the absolute values of standardized coefficients for all of the features decreased when the users ranked in the top with regard to follower increase were targeted. More importantly, as shown in Fig. 5a, the standardized coefficient of the related rate surpassed initial followers for those users ranked in the top 1% in follower increase. Second, we trained multiple random forest models on users with different follower increase rankings and compared the features' relative importance in the different models (Fig. 5b). Fig. 5b illustrates that the importance of initial follower decreases for users with higher rankings of follower increases. On the other hand, the related rate became more

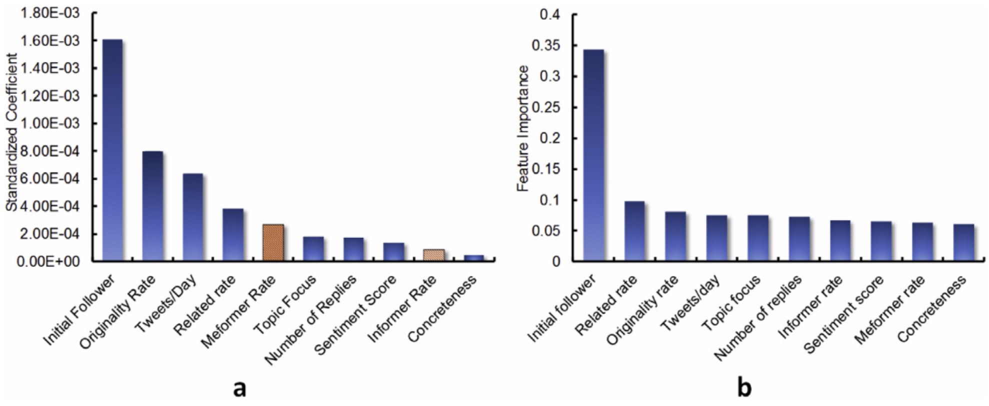

Fig. 3. a, Standardized coefficient for the negative binomial regression model ranked by their absolute value. Solid and striped bars indicate positive and negative values, respectively. **b**, Feature weights of the random forest regression model.

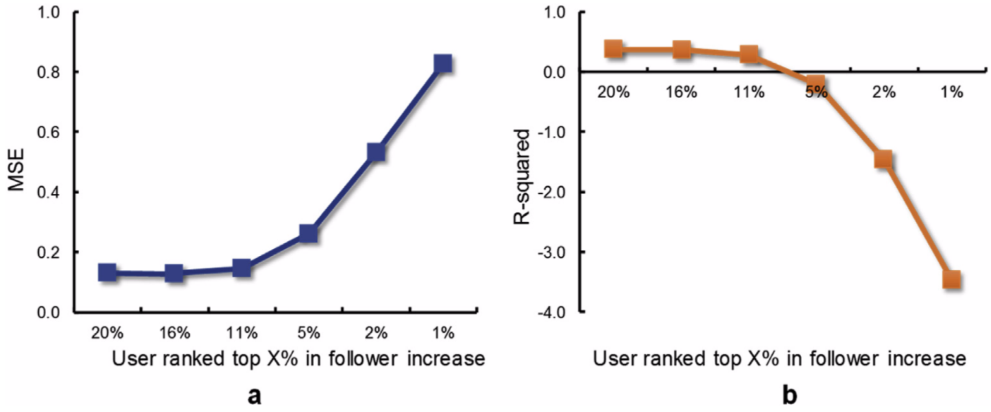

Fig. 4. MSE and R-squared value of random forest regression model in step 1 for users with different ranges of follower increase.

significant and its importance was very close to initial follower (0.165 vs. 0.170) for the users with the top 1% in follower increase. The relative importance of other features did not change significantly. These findings suggested that the related rate of top users contributed to their high follower increase (in comparison to ordinary users for whom the initial follower was the most significant feature). These findings also encouraged us to further explore the emerging influencers among the top 1% of users concerning follower increase and to seek a better understanding of their uniqueness about features that contributed to it.

Fig. 5c illustrates the performance of random forest models for users with different levels of follower increase. The figure shows that the numbers of follower increase for users with a higher number of follower increases were difficult to ascertain using a random forest regression. To investigate the influence of the size of the training dataset, we also subsampled six datasets from users of top 20%, 16%, 11%, 5%, 2%, and 1% with the same size (692 users, which is the number of users after filtering who ranked top 1% in follower increase) to train and cross-validate the random forest models. To mitigate the influence of random subsampling, we repeated the test 10 times and looked at the overall distribution of the results. The result showed that, even with the same amount of training sample, the performance of models decreases as the follower-increase ranking of the users in the training dataset increases. This decrease confirmed that the increases in followers is more difficult to predict for the top emerging influential users than for ordinary users and indicated the existence of multiple patterns in follower increase for the top users.

3.3. Step 3: characterizing the contributors to public information and warning within the emerging influential users

Step 3 permitted characterization of the emerging influential users through verifying that the unpredictability was caused by the multiple patterns that generated increases in followers, such as the posting of quality information related to the disaster or large numbers of initial followers. This step focused on the top 1% of uses in follower increase for two reasons. First, these users are typical examples of emerging influential users considering their significant follower gain during Harvey. Second, they have different patterns with the general public in terms of the relationship between follower increase and the features, as is shown in previous steps.

3.3.1. Clustering the emerging influencers for identifying multiple follower-increase patterns

To separate the top influential users who emerged during Hurricane Harvey, we used a weighted clustering algorithm to divide the top 1% of the users by their content and behavioral features. The weights of the features were calculated as their relative importance from the trained random forest model for users in the top 1% increase in followers. Using

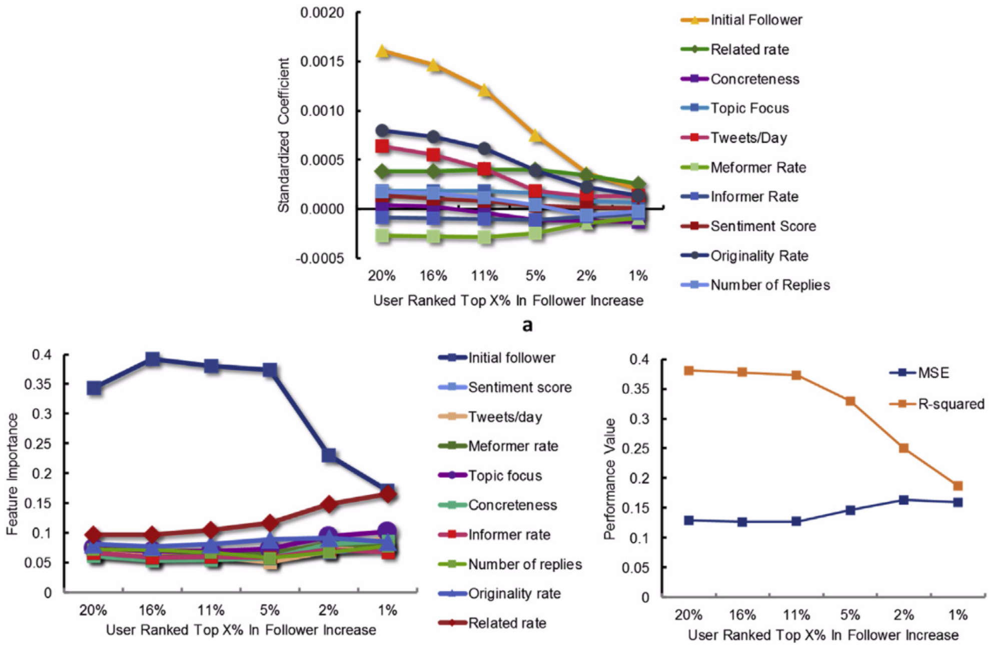

Fig. 5. Results of training regression models with users with different ranges of follower increase. **a**, Change of the standardized coefficient of each feature in negative binomial regression models trained with users of different ranges of follower increase (points marked as triangle means that these features failed to reject the null hypothesis of not having a significant effect on the corresponding models with the p-value > 0.05). **b-c**, Change of feature importance and performance in random forest regression models trained with users with different ranges of follower increases.

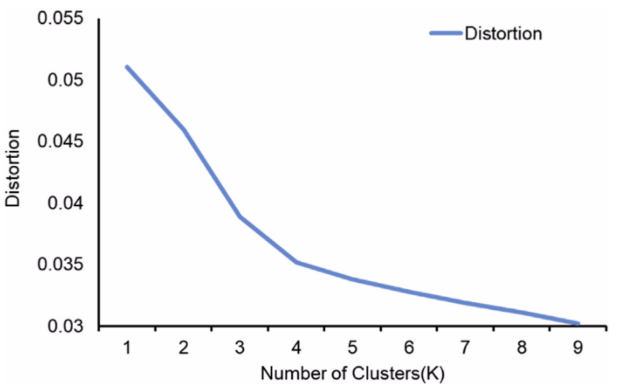

Fig. 6. The distortion of clustered results from running K-Means clustering with a different number of clusters (K) specified. Noted that the "Elbow" is identified to be at K = 4.

the elbow method [35] shown in Fig. 6, we determined that K = 4 was optimum.

To identify the most significant features for clustering the emerging influential users and visualize them in 3-d space, we processed Principal Component Analysis on the features of the top 1% Houstonian Twitter users in follower increase. To find how many components are required to represent the dataset, we calculated the percentage of variance explained for all the principal components. The first three components combined explain over 78.75% of the total variance, with the percentage of variance explained drastically decreasing for the following components. As a result, we decided to use the first three components as our principal components. Among weights of features in the three principal components, initial follower, related rate, and original rate stand out compared to other features regarding PCA component weights. To compare the three features to the three main principal components determined above, we also calculated the ratio of variance explained for each feature. The results show that the features of initial follower, related rate, and original rate combined have over 76.53% of variance explained, while all other features are having a drastically lower percentage of variance explained. This observation shows that these three features have a significant contribution to distinguishing the data points when separating the top 1% of Houstonian Twitter users in follower increase into different categories.

Fig. 7a illustrates the top 1% of users in a three-dimensional space, with the x, y, and z-axes representing an initial follower, related rate, and originality rate, respectively. The purple cluster with the square marks is the cluster of users that exhibited high related rates and low to moderate levels of initial followers. The green cluster with circle marker had a high number of initial followers and low to moderate levels of related rates. The gray cluster with triangle marker had high originality rates. Finally, the yellow cluster with x-shaped marker had low to moderate related rates, initial followers, and originality rates. Fig. 7b shows that users in the purple cluster with square marker had the most significant mean and median follower increases among the users in the top 1% of follower increases.

Fig. 8 shows the word clouds of the tweets posted by users in the four clusters. To better highlight the difference between the tweet content of the users in different clusters, words with high frequency in all four clusters (i.e., "Houston" and "thanks") are removed from the word clouds. The users in the purple (square) cluster posted disaster situational awareness information (shown in Fig. 8a), while others emphasized on emotional or comforting posts during Hurricane Harvey (shown in Fig. 8b, c, and d). Based on these results, the authors argue that, among the emerging influential users, people who continue to post disaster-related information would gain the most significant follower increases because of their contributions to public information and warning during disasters. The analysis also demonstrated that the contributors to public information and warning (i.e. users in the purple/square cluster) could have either high or low originality rates, suggesting that contributors play a variety of roles. Contributors with high originality rates could be 'information generators' who provide situational disaster or response information based on their professional knowledge or the resources available to them. In contrast, contributors with low originality rates would serve primarily as 'information disseminators'; they could collect and distribute useful information during disasters.

3.3.2. *Characterizing the contributors to public information and warning*

The analysis raises an interesting question: who are the emerging influential users who made the greatest contributions to public information and warning during Hurricane Harvey? Specifically, who were the Twitter users in the purple cluster with square marker in Fig. 7? To examine the characteristics of these contributors, we annotated the 179 users in the cluster with the highest related rates according to the industry of the account users, verified or not, and whether or not they were individual or organizational accounts. Fig. 9a shows that most account users worked for the media industry, followed by users from the public sector and then those from research organizations. Accounts from the media industry included, but were not limited to, official accounts from TV channels, news reporters, journalists, producers of weather reports, and participatory media producers. Accounts from the public sector include politicians, as well as lawyers, and other professionals working in disaster management agencies. Fig. 9b shows that more than half (55%) of the influential accounts were verified; verification of an account contributed to an increase in followers. Fig. 9c illustrates that only 44 (25%) accounts were organizational accounts; the remaining 75% of all accounts belonged to individual users. In summary, the clustering analysis for the users who were ranked in the top 1% in follower increase showed that professionals made great contributions to public information sharing and warnings during Hurricane Harvey. They gained significant follower increases with relatively low initial follower numbers because

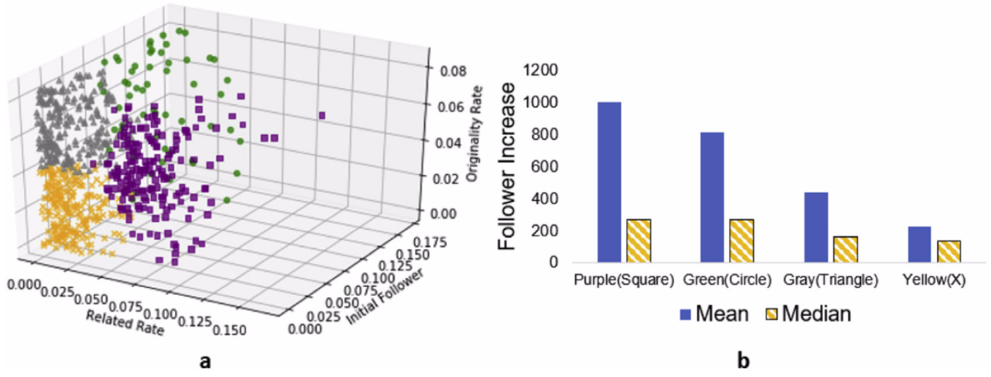

Fig. 7. The result of separating the top 1% of users in follower increase into different clusters. **a,** 3d space of emerging influential users with axes being the initial follower number (X), related rate (Y), and originality rate (Z). The purple cluster with square marker is the cluster of emerging influential individuals with the highest related rate. **b,** A comparison of the follower increase of different clusters by mean and median. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

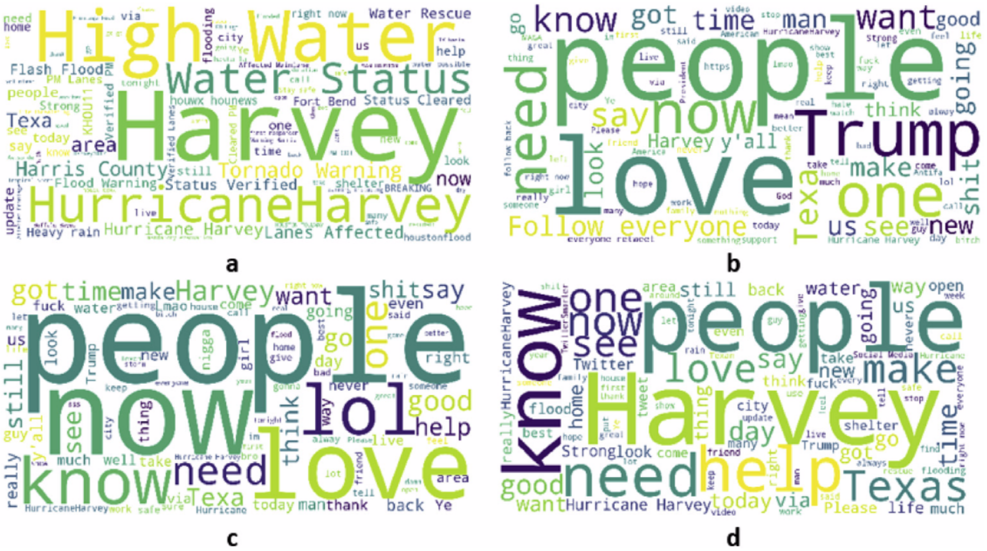

Fig. 8. Word cloud of tweets by users from the purple (square), green (circle), gray (triangle), and yellow (cross) cluster in Fig. 8, respectively. Panel **a** represents the word cloud of the cluster of emerging influential individuals who contributed to public information and warnings. Panel **b**, **c**, and **d** represent the word cloud of the emerging influential users who expressed their feelings but provided limited information about the hurricane. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

they used their professional knowledge and consistently posted information on the disaster situation.

4. Discussion

4.1. Summary of findings

The findings of this study identified key features that shaped the emergence of influential users who disseminated information during disasters. First, for the average Twitter user, the increase in followers was positively related to the initial number of followers, the disaster-related rate, the number of tweets posted and replies, and the originality rate. Tweets about the user him/herself were negatively related to an increase in followers. The general pattern of Twitter follower increases (shown in Section 3.1 and Fig. 4) during Hurricane Harvey suggested that the public tended to follow users that posted objective information related to the disaster and posted consistently. Additionally, the initial number of followers was the most significant predictor of follower increases for ordinary users in both the negative binomial regression and random forest regression models. This observation suggested that users who already enjoyed a high number of followers could benefit from a 'Matthew effect' (advantage begets further advantage, [38,39]) for greater gains in followers.

Second, step 2 revealed that, when targeting the users who ranked highest (i.e., the top 1–2%) in follower increases, the disaster-related rate represented the most important feature that produced a greater following; high numbers of initial followers proved less relevant for these users. Furthermore, step 3 confirmed that users who posted a high proportion of disaster-related information gained the most followers among people who ranked in the top 1% of follower increases. These results distinguished the contributor to public information and warning from the Twitter users who gained followers due to other reasons, such as a large number of initial followers. This finding confirmed the presence of a self-organizing feature of social media during disasters; the public recognized the people who contributed to public information dissemination. Furthermore, this finding confirmed that user behavior contributed significantly to the dynamics of network structures. People who consistently produced high-quality information about the disaster (i.e., users in the purple (square) cluster in Fig. 8) defied the Matthew effect.

Third, the annotation analysis in step 3 demonstrated that the emerging influencers who contributed to the dissemination of information during Hurricane Harvey were the professionals who worked in the media industry, public sector, and research organizations. During step 3 it became clear that most of the information hubs that contributed to public information and warning during Harvey were

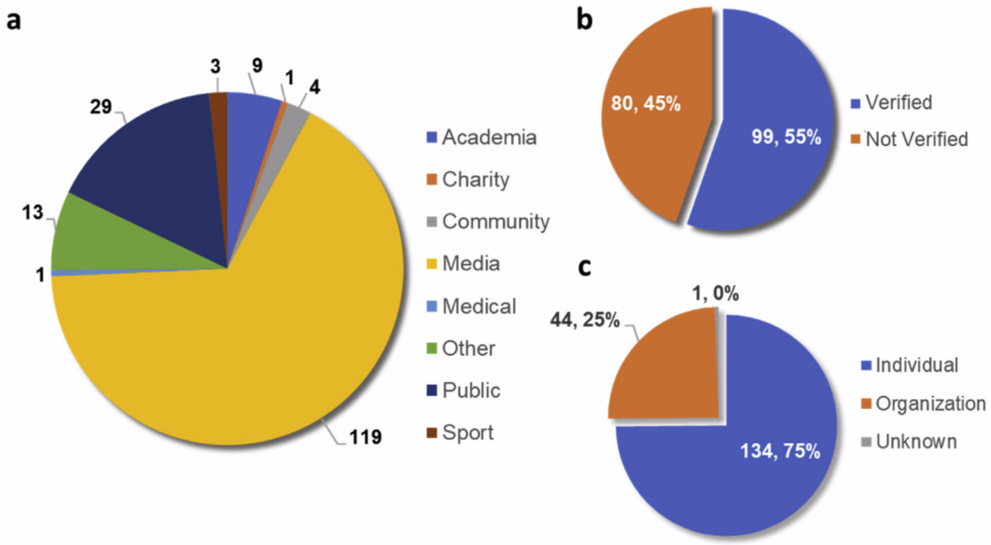

Fig. 9. Statistics of emerging influential users who made a significant contribution to public information and warning during Hurricane Harvey. **a**, the distribution of industries with which the emerging influential accounts were affiliated. **b**, the number of emerging influential accounts that were verified or unverified. **c**, The number of individual accounts versus organizational accounts.

comprised of individuals rather than organizational accounts. Yet, disaster studies often fail to examine this sector of users. Existing research about social media usage during disasters have recognized the significant contributions of organizational users, including those from government organizations (GOs), non-governmental organizations (NGOs), and mass media agents [40–42]. For example, according to case studies on Typhoon Haiyan [41] and Hurricane Sandy [42], most tweets posted by newsagents or government organizations were official disaster information updates. However, the general public primarily uses social media to express their emotional reactions and share personal information during disasters [40]. This study identified the contribution of professional individuals (e.g., journalists, news reporters, and meteorologists) during disasters. Together with other organizations, these users represented important nodes in the networks of information disseminated during disasters.

4.2. The theoretical contribution and practical value

This study was the first to focus on an examination of follower increases on Twitter during a disaster. The theoretical contribution of this study on the social network and human behavior is two-fold. First, this study took an initial step toward an exploration of the ways and extent to which Twitter usage behaviors could influence a network structure during extreme events such as natural disasters. During disasters, contributors to public information and warning could challenge the Mathew effect that suggests that the number of initial followers would determine the influence gained on social networks. Rather, a tendency toward self-organization prevailed in the dissemination of disaster information and people were more likely to secure access to needed disaster-related information through other social media users (i.e. 'following'). Future studies could focus on the production of automation tools with guidelines for social media users that could enhance and inform self-organized communication on social media platforms to assure reliable public information and warning during disasters [43]. Second, this study also illuminated the roles and functions of different types of information hubs during disasters. Understanding how disaster situation information was created by "information generators" and disseminated by "information distributors" based on network modeling and graph theory could help to further improve the efficiency and effectiveness of public information sharing and warnings during such events.

The practical value of this study also proved to be two-fold. First, this study provided guidelines for people who would like to contribute to disaster information dissemination; they should post disaster-related information in clear language, take care to avoid irrelevant information, and abstain from posting personal views or information. Second, this study demonstrated that professional individuals—not only organizational users—played a significant role in disaster information dissemination. Disaster management agencies could include professional individuals in both the planning and processing of public information sharing and warnings during disasters in the future.

4.3. Limitation

The main limitation of this study is that it did not permit a detailed exploration of the causal relationship between the tweeting activity of each emerging influencer and the events of the people following the emerging influencer. This relationship would help to explain the interaction among the dynamics of the disaster situation, information dissemination, and social network evolution. This limitation was due primarily to the unavailability of the exact timestamp of following/unfollowing activity during Hurricane Harvey (i.e., A followed/unfollowed B at time X). Unfortunately, such data no longer exist through the Twitter API. During hurricane seasons, researchers could begin to monitor the follower lists of the Twitter users who are based in the potentially impacted areas before the hurricane lands.

5. Conclusion

This study focuses on the emergence of influential individual users on Twitter and explores the reason why they achieved a significant number of followers during one of the most devastating disasters: Hurricane Harvey. The diverse patterns of increases in followers for the emerging influencers may be separated into several types. Among those types, Twitter users who posted objective disaster-related information in clear language and a consistent fashion tended to show an increase in followers and emerged as influential users. Additionally, professionals (from the media industry, public agencies and departments, and research organizations) made significant contributions to public information and warning during the hurricane. The findings highlight the common attributes of emerging influential social media users and the crucial role they play in the self-organized dissemination of disaster information.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant Number IIS-1759537 and the Amazon Web Services (AWS) Machine Learning Award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation and Amazon Web Services (AWS).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijdrr.2019.101204.

References

[1] Irina Shklovski, Moira Burke, Sara Kiesler, Robert Kraut, Technology adoption and use in the aftermath of hurricane katrina in New Orleans, Am. Behav. Sci. 53 (8) (2010) 1228–1246 https://doi.org/10.1177/0002764209356252.

[2] U.S. Department of Homeland Security, National incident management system, https://www.fema.gov/pdf/emergency/nims/NIMS_core.pdf, (2008).

[3] Nicki Dabner, 'Breaking ground' in the use of social media: a case study of a university earthquake response to inform educational design with facebook, Internet High Educ. 15 (1) (2012) 69–78 https://doi.org/10.1016/j.iheduc.2011.06.001.

[4] Shampy Kamboj, Bijoylaxmi Sarmah, Shivam Gupta, Yogesh Dwivedi, Examining branding Co-creation in brand communities on social media: applying the paradigm of stimulus-organism-response, Int. J. Inf. Manag. 39 (April) (2018) 169–185 https://doi.org/10.1016/j.ijinfomgt.2017.12.001.

[5] Serena Tagliacozzo, Michele Magni, Government to citizens (G2C) communication and use of social media in the post-disaster reconstruction phase, Environ. Hazards 17 (1) (2018) 1–20 https://doi.org/10.1080/17477891.2017.1339012.

[6] Adrien Guille, Hakim Hacid, Cecile Favre, Djamel a Zighed, Information diffusion in online social networks: a survey, SIGMOD Rec. 42 (2) (2013) 17–28 https://doi.org/10.1145/2503792.2503797.

[7] Daniel M. Romero, Brendan Meeder, Jon Kleinberg, Differences in the mechanics of information diffusion across topics, Proceedings of the 20th International Conference on World Wide Web - WWW '11, 695, 2011 https://doi.org/10.1145/1963405.1963503.

[8] Seth A. Myers, Leskovec Jure, The bursty dynamics of the twitter information network, Proceedings of the 23rd International Conference on World Wide Web - WWW '14, 913–24, ACM Press, New York, New York, USA, 2014, https://doi.org/10.1145/2566486.2568043.

[9] Fabián Riquelme, Pablo González-Cantergiani, Measuring user influence on twitter: a survey, Inf. Process. Manag. 52 (5) (2016) 949–975 https://doi.org/10.1016/j.ipm.2016.04.003.

[10] abc 13. n.d. "Mayor Turner Proclaims May 2 as 'Jeff Lindner Day' in Houston after Famed Meteorologist." Accessed May 3, 2018. https://abc13.com/society/mayor-turner-proclaims-may-2-as-jeff-lindner-day-in-houston/3426127/.

[11] Ravi Kumar, Jasmine Novak, Andrew Tomkins, Structure and evolution of online social networks, Link Mining: Models, Algorithms, and Application, Springer New York, New York, NY, 2010, pp. 337–357 https://doi.org/10.1007/978-1-4419-6515-8_13 s.

[12] Jure Leskovec, Jon Kleinberg, Christos Faloutsos, Graphs over time, Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining - KDD'05, vol. 177, ACM Press, New York, New York, USA, 2005, https://doi.org/10.1145/1081870.1081893.

[13] Peng Wang, Bao Wen Xu, Yu Rong Wu, Xiao Yu Zhou, Link prediction in social networks: the state-of-the-art, Sci. China Inf. Sci. 58 (1) (2014) 1–38 https://doi.

org/10.1007/s11432-014-5237-y.

[14] Lars Backstrom, Jure Leskovec, Supervised random walks: predicting and recommending links in social networks, In Proceedings of the Fourth ACM International Conference on Web Search and Data Mining - WSDM, vol. 11, ACM Press, New York, New York, USA, 2011, p. 635 https://doi.org/10.1145/1935826.1935914.

[15] Yongli Li, Peng Luo, Zhi-ping Fan, Kun Chen, Jiaguo Liu, A utility-based link prediction method in social networks, Eur. J. Oper. Res. 260 (2) (2017) 693–705 https://doi.org/10.1016/j.ejor.2016.12.041.

[16] Dong Li, Yongchao Zhang, Zhiming Xu, Dianhui Chu, Sheng Li, Exploiting information diffusion feature for link prediction in sina weibo, Sci. Rep. 6 (August 2015) (2016) 1–8 https://doi.org/10.1038/srep20058.

[17] Yadong Zhou, Beibei Zhang, Xiaoxiao Sun, Qinghua Zheng, Ting Liu, Analyzing and modeling dynamics of information diffusion in microblogging social network, J. Netw. Comput. Appl. 86 (May) (2017) 92–102 https://doi.org/10.1016/j.jnca.2016.09.011.

[18] Demetris Antoniades, Constantine Dovrolis, Co-evolutionary dynamics in social networks: a case study of twitter, Proceedings - 10th International Conference on Signal-Image Technology and Internet-Based Systems, vol. 2014, SITIS, 2015, pp. 361–368 https://doi.org/10.1109/SITIS.2014.68.

[19] James P. Gleeson, Kevin P. O'Sullivan, Raquel A. Baños, Yamir Moreno, Effects of network structure, competition and memory time on social spreading phenomena, Phys. Rev. X 6 (2) (2016) 1–22 https://doi.org/10.1103/PhysRevX.6.021019.

[20] Nicola Barbieri, Francesco Bonchi, Giuseppe Manco, Who to follow and why, Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '14, ACM Press, New York, New York, USA, 2014, pp. 1266–1275 https://doi.org/10.1145/2623330.2623733.

[21] Zhepeng (Lionel) Li, Xiao Fang, R. Olivia, Sheng Liu, A survey of link recommendation for social networks, ACM Trans. Manag. Inf. Syst. 9 (1) (2017) 1–26 https://doi.org/10.1145/3131782.

[22] C.J. Hutto, Sarita Yardi, Eric Gilbert, A longitudinal study of follow predictors on twitter, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI, vol. 13, ACM Press, New York, New York, USA, 2013, p. 821 https://doi.org/10.1145/2470654.2470771.

[23] NATIONAL HURRICANE CENTER, Costliest, U.S. Tropical Cyclones Tables Updated, 2018.

[24] Time, Hurricane Harvey: the U.S.'s First Social Media Storm vol. 2017, (2017).

[25] Jacob Goldenberg, Sangman Han, Donald R. Lehmann, Jae Weon Hong, The role of hubs in the adoption process, J. Mark. 73 (2) (2009) 1–13 https://doi.org/10.1509/jmkg.73.2.1.

[26] Steven Bird, Ewan Klein, Edward Loper, Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, " O'Reilly Media, Inc.", 2009.

[27] Virginia Teller, Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition daniel Jurafsky and James H. Martin (university of Colorado, boulder) upper saddle river, NJ: prentice Hall (prentice Hall Se, Comput. Linguist. 26 (4) (2000) 638–641 https://doi.org/10.1162/089120100750105975.

[28] Marc Brysbaert, Warriner Amy Beth, Victor Kuperman, Concreteness ratings for 40 thousand generally known English word lemmas, Behav. Res. Methods 46 (3) (2014) 904–911 https://doi.org/10.3758/s13428-013-0403-5.

[29] C J Hutto Eric Gilbert, Vader: a parsimonious rule-based model for sentiment analysis of social media text, Eighth International Conference on Weblogs and Social Media (ICWSM-14), 2014 http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf Available at: (20/04/16).

[30] Joseph M. Hilbe, Negative Binomial Regression, Cambridge University Press, 2011.

[31] Robin Genuer, Jean-michel Poggi, Christine Tuleau-malot, Variable selection using random forests, Pattern Recogn. Lett. 31 (14) (2010) 2225–2236 https://doi.org/10.1016/j.patrec.2010.03.014.

[32] Ulrike Grömping, Ulrike G. Römping, "Variable importance assessment in Regression : linear regression versus random forest variable importance assessment in Regression : linear regression versus random forest, Am. Statistician 1305 (2012), https://doi.org/10.1198/tast.2009.08199.

[33] Leo Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al., Scikit-learn: machine learning in {P}ython, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[35] Robert L. Thorndike, Who belongs in the family? Psychometrika 18 (4) (1953) 267–276 https://doi.org/10.1007/BF02289263.

[36] J.A. Hartigan, M.A. Wong, Algorithm as 136: a K-means clustering algorithm, Appl. Stat. 28 (1) (1979) 100 https://doi.org/10.2307/2346830.

[37] Jacob Cohen, Statistical Power Analysis for the Behavioral Sciences, second ed., Erlbaum Associates, Hillsdale, 1988.

[38] Wesley R. Hartmann, Puneet Manchanda, Harikesh Nair, Matthew Bothner, Peter Dodds, David Godes, Kartik Hosanagar, Catherine Tucker, Modeling social interactions: identification, empirical methods and policy implications, Mark. Lett. 19 (3–4) (2008) 287–304 https://doi.org/10.1007/s11002-008-9048-z.

[39] Daniel Rigney, The Matthew Effect: How Advantage Begets Further Advantage, Columbia University Press, 2010.

[40] Clarissa C. David, Corpus Ong Jonathan, Erika Fille T. Legara, Tweeting supertyphoon Haiyan: evolving functions of twitter during and after a disaster event, PLoS One 11 (3) (2016) 1–19 https://doi.org/10.1371/journal.pone.0150190.

[41] Bruno Takahashi, Edson C. Tandoc, Christine Carmichael, Communicating on twitter during a disaster: an analysis of tweets during Typhoon Haiyan in the Philippines, Comput. Hum. Behav. 50 (2015) 392–398 https://doi.org/10.1016/j.chb.2015.04.020.

[42] Bairong Wang, Jun Zhuang, Crisis information distribution on twitter: a content analysis of tweets during hurricane Sandy, Nat. Hazards 89 (1) (2017) 161–181 https://doi.org/10.1007/s11069-017-2960-x.

[43] Cheng Zhang, Chao Fan, Wenlin Yao, Xia Hu, Ali Mostafavi, Social media for intelligent public information and warning in disasters: an interdisciplinary review, Int. J. Inf. Manag. 49 (December) (2019) 190–207 https://doi.org/10.1016/j.ijinfomgt.2019.04.004.

[44] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, Charles J. Stone, Classification and Regression Trees, (2017) Routledge. https://doi.org/10.1201/9781315139470.