Network Analysis of Tweets from the Nebraska Flood Extreme Weather Event

Hannah Ker

June 2019

This report describes the methodology and key findings from a social network analysis of Tweets relating to the Nebraska flooding that occurred in March, 2019. The approach described in this report is a pilot analysis that aims to explore a potential methodology for identifying the primary news sources on social media during an extreme weather event.

SECTION 1: Methods

This section describes the methods used in the network analysis of tweets from the Nebraska Flood weather event. These methods were used to identify key news sources on Twitter during this weather event.

1.1: Data Collection

The midwestern United States experienced significant flooding, beginning in mid-March 2019. Flooding primarily occurred along the Missouri River and was caused by sudden rains, combined with temperature increases, leading to significant snow melt[1]. According to a New York Times article[2] at least two people in Nebraska had died and the governors of Nebraska, South Dakota, and Wisconsin had declared emergencies at the time of writing on March 18, 2019.

Beginning on March 22, 2019, we collected tweets relating to the flooding in Nebraska in real-time using the Twitter API. Tweets were collected by filtering according to hashtags. To ensure that the tweets in our datasets were relevant to the flood event, we only collected tweets that included at least one of the following hashtags:

- #nebraskaflood
- #flood2019
- #nebraskastrong
- #missouririver
- #nebraskaflood2019
- #prayfornebraska

These hashtags were identified by manually surveying tweets that were deemed relevant to the flood and noting commonly-used hashtags. The Socioviz hashtag co-occurrence tool[3] was also used to identify several commonly-occurring hashtags.

Tweets relating to the Nebraska Flood event were streamed from March 22 to March 26, resulting in a total of 12,827 tweets.

[1] https://www.nytimes.com/2019/03/18/us/nebraska-flooding-facts.html
[2] ibid.
[3] http://socioviz.net/SNA/eu/sna/login.jsp

1.2: Building Network Structure

Following data collection, the dataset of tweets was converted into a network structure of nodes and edges. Nodes are vertices in the network, connected by edges that are, in this case, directional.

This network is intended to model flows of information between Twitter users during the Nebraska flood. As such, nodes in this network correspond to Twitter user accounts. One user account corresponds to one node in the network. The edges connecting these nodes are informational exchanges between accounts, as represented by mentions, retweets, and replies. As information exchange is directional (ie. it has a source, the actor delivering the information, and a target, the actor receiving the information), the edges in this social network are also directional. Edges point from source to target. For each type of interaction between Twitter user accounts, directionality is as follows:

- Retweets: Source = Retweeted; Target = Retweeter
- Mentions: Source = Mentioned; Target = Mentioner
- Replies: Source = Replied to; Target = Replier

Edges may also be weighted according to the strength of interactions between user accounts. For example, if User A retweets User B three times, then the edge connecting these nodes would have a weight of three. Figure 1.1 exemplifies this approach to modeling a social network of Twitter interactions. Nodes A, B, C, and D represent Twitter accounts and the edges between these nodes represent retweets, mentions, and replies.

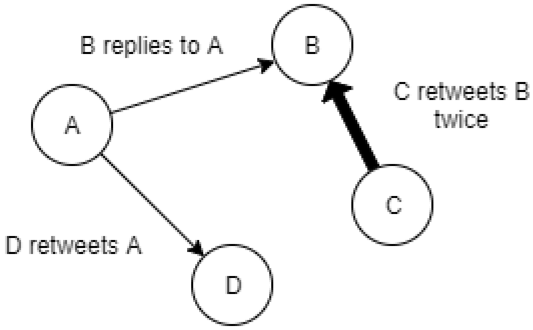


Figure 1.1: Example network

This approach to modeling a social network of information flow using Twitter data follows past work in this domain. Wallis, Fisher, and Lvov (2015) investigate the network structure of communication on Twitter during the 2012 London Olympics, similarly modeling network nodes as Twitter accounts and edges as replies, retweets, and mentions between accounts. Yang and Counts (2010) analyse patterns in the speed, scale, and range of information diffusion on Twitter by creating a network with nodes as Twitter accounts and edges as mentions between accounts. Additionally, Chatfield and Brajawidagda (2012) created a network of retweets between Twitter accounts to determine the pattern of information diffusion during a tsunami.

The resulting social network generated using this approach has a total of 8103 nodes, connected by 9702 edges. 8585 of these edges are retweets, 914 edges are mentions and 203 edges are replies.

1.3: Basic Network Analysis

The freely-available software tool, Gephi[4], was used to visualize and analyse the social network of information flow during the Nebraska flood event. Basic measures; such as average path length, and the network diameter; were calculated to quantitatively characterize the structure of this social network. These measures offer a means to compare various features of this network with other networks.

1.4: Analysis of Node Centrality

Node centrality is a commonly studied feature of social networks and consists of a number of properties that capture the prominence of a node within its network (Borgatti et al., 2009). In this case, the relative centrality of a Twitter account (node) within the network is considered to be indicative of that account's influence over the flow of information within the network. An account that is more central in the network is considered to be a more influential source of information relating to the Nebraska flood event.

The centrality of a node within a network can be measured in a variety of ways. The following is an overview of common measures of node centrality. Note that each of these measures are applied to individual nodes, rather than the network as a whole.

Degree centrality: The simplest measure of node centrality, degree centrality is a measure of the number of "one-hop" connections (edges) between nodes in a network. For example, a node with a degree of three would have three direct connections to other nodes. For directional networks, degree centrality can also be measured in terms of in-degree (incoming connections) and out-degree (outgoing connections). A node with a high degree centrality is considered to be popular and well-connected to other nodes in a network. However, degree centrality does not take into account the relative popularity of a node's connections. For example, a node that is connected to many other well-connected nodes may be more central than another node that is connected to the same number of less-popular nodes.

Betweenness centrality: Betweenness centrality measures the number of times that a node lies on the shortest path between two other nodes. As such, nodes with a high betweenness centrality may be considered as "bridges" or informational/knowledge brokers between communities in a network.

PageRank: PageRank is an algorithm used by the Google Search Engine to rank the importance of websites (Page et al., 1999). This algorithm measures the importance of node A in a network (eg. website on the web) by taking into account both the degree centrality of node A and the degree centrality of all the nodes that are connected to node A. This algorithm has been applied to Twitter networks to identify central nodes (accounts), such as in the case of Willis, Fisher, and Lvov (2015).

[4] https://gephi.org/

In the case of social networks of Twitter interactions, past research efforts do not use a consistent centrality measure when identifying the most central nodes in the network. As such, we applied all three measures identified above and looked for nodes that ranked highly according to multiple measures.

Phase 1: Analysis of all nodes in network

The degree centrality, betweenness centrality, and PageRank values were calculated for all nodes in the network. The top ten nodes with the highest values for each centrality measure were reported. The nodes that appeared in at least two out of the three top ten rankings were considered to be among the most central nodes in the network. For each of these nodes, the attributes in Table 1.1 were reported. These attributes are intended to characterize the most central nodes in the network.

Table 1.1: Node attributes

Attribute	**Description**	**Potential values**
Twitter handle	Account username	N/A
Number of followers	Number of Twitter accounts following the account	N/A
Type of account	The type of entity that the Twitter account represents	Personal Commercial Government Other
Geographic scope	Refers to the geographic scope of location-based information that is being tweeted. This attribute applies if the Tweet content associated with the account is location-specific (eg. communicates information about a specific location – determined through manual review of a sample of the account's latest ten tweets)	Local Regional National International
Associated location	The location tagged in the account's profile	N/A

Phase 2: Analysis of nodes from largest community

Following Phase 1, the entire network was broken down into communities using Blondel et al's (2008) modularity class algorithm (implemented in Gephi). Exemplified by Figure 1.2, this algorithm identifies groups of densely-connected nodes in a network that interact with each other more frequently than with nodes from other communities (Blondel et al, 2008).

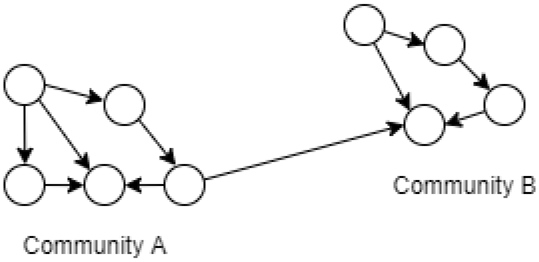


Figure 1.2: Example of communities in a hypothetical network

It is hypothesized that the Tweets and interactions from the largest community of nodes will be more directly related to the Nebraska flood event than those from smaller, more peripheral communities. The peripheral communities may contain more noise and be less directly related to the Nebraska flood event. Accordingly, the largest community of nodes was selected and the analysis from Phase 1 was repeated on this subset of the network. Following the hypothesis identified above, it is assumed that the most central nodes from this subset of the network will be more indicative of the key news sources relating to the Nebraska flood than the most central nodes from the entire network.

SECTION 2: Results

This section presents and interprets the results of the social network analysis of tweets.

2.1: Structure of Social Network

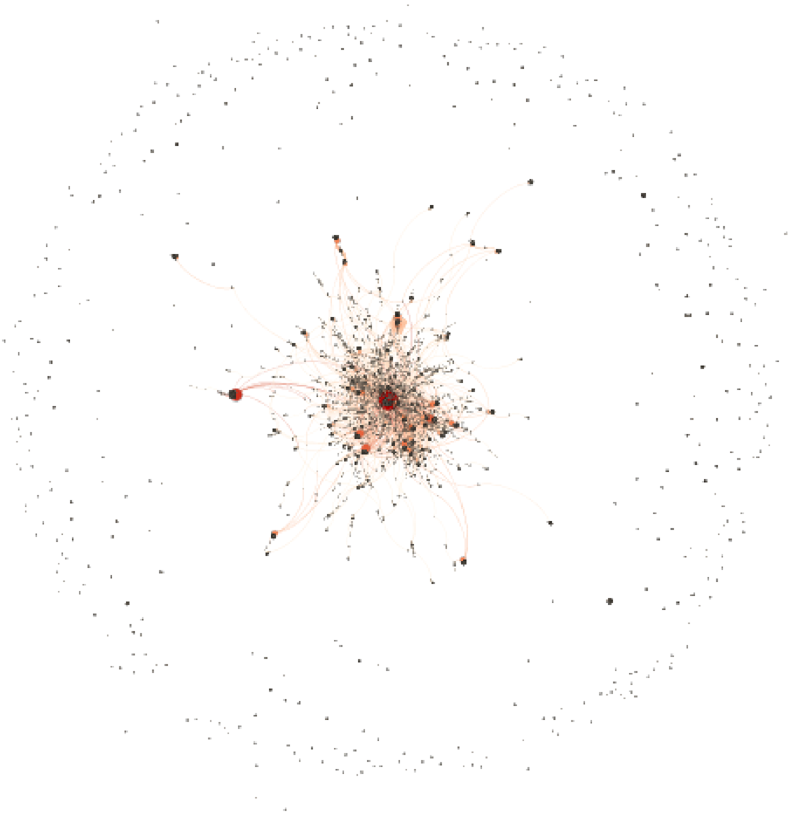
Figure 2.1 illustrates the structure of the social network of Tweets, modeled according to the framework identified in Figure 1.1. The colour intensity and the size of each node corresponds to its degree. Upon visual inspection of this image, one can see that the network has many small peripheral clusters that are disconnected from each other. The center of the network contains a much larger cluster of more densely connected nodes.

Table 2.1: Structural characteristics of social network

Metric	**Description**	**Value**
Average path length	Average distance between two nodes[5]	5.613
Graph density	Measure of network “completeness” – ie. the extent to which all possible connections between nodes exist[6]	0.00
Network diameter	Longest path between two nodes[7]	13
Average degree	Average number of connections for each node	1.197

[5] https://github.com/gephi/gephi/wiki/Average-Path-Length
[6] https://github.com/gephi/gephi/wiki/Graph-Density
[7] https://github.com/gephi/gephi/wiki/Diameter

Figure 2.1: Visualization of social network



Note: a low graph density and average degree of 1.197 indicates that the nodes in this network are sparsely connected. On average, each node is only connected to approximately one other node.

2.2: Analysis of Node Centrality

Phase 1: Analysis of all nodes in network

Table 2.2 shows the most central nodes in the network, according to degree centrality, betweenness centrality, and the PageRank algorithm. Accounts highlighted in blue are those that have top scores (rank in the top ten) for all three centrality measures and accounts highlighted in green are those that have top scores for two of the three centrality measures.

The Twitter account for Governor Pete Ricketts of Nebraska ranks the highest across all three measures of node centrality for the entire social network. This result indicates that the Tweets from the Governor's account have the most direct interactions with other Twitter accounts (through mentions, retweets, and replies) than any other account. Most of the direct connections to Governor Ricketts' account are outbound, indicating that other accounts are interacting with his Tweets, rather than the inverse. Governor Ricketts' account also most frequently lies on the shortest path between any other two accounts in the network, indicating that information shared on his account may reach many communities of Twitter users. The high PageRank score also indicates that other highly central accounts are interacting with Governor Ricketts' account.

Governor Ricketts' account is the only Twitter account (ie. node in the network) to rank in the top ten across all three measures of node centrality. There are six accounts appear twice in this ranking. Interestingly, the account "DuffelBagDustin" appears to be an outlier due to a low number of followers. The account "Barbi_Twins" is also surprisingly based in Malibu, California (although this is not necessarily where the account's user was Tweeting from. The majority of central accounts in Table 2.3 are personal, rather than official government accounts or commercial news sources. This finding may indicate that there was a significant unofficial news presence during the Nebraska flood, with many users interacting with information in Tweets from these unofficial sources. This finding may also suggest a certain degree of noise in the dataset of Tweets, as there is the potential that some of the Twitter accounts identified in Table 2.3 were not tweeting about news relating to the flood (and perhaps instead shared a notable video or picture relating to the flood).

Table 2.3 summarizes the characteristics of the nodes that appear in at least two out of the three top ten centrality rankings (those coloured in green and blue in Table 2.2). Each of the Twitter accounts listed in Table 2.3 may be understood as a significant source of information during the Nebraska flood event.

Table 2.2: Most central nodes in the network, according to three measures

	Degree	**Betweenness**	**PageRank**
1	GovRicketts	GovRicketts	GovRicketts
2	UniqueAdInc	NETAGBohac	NEMAtweets
3	Matt_Davison	_ashmuell	USDA
4	SarahFiliKETV	NebToday	SecretarySonny
5	DuffelBagDustin	ExtensionBen	Matt_Davison
6	RBrex34	NEMAtweets	SarahFiliKETV
7	cucoachmac	UNLExtension	DuffelBagDustin
8	ChinhDoan	NENationalGuard	jtimberlake
9	Barbi_Twins	NEStatePatrol	NENationalGuard
10	MistaBRONCO	UNL_CropWatch	Barbi_Twins

Table 2.3: Characteristics of most central nodes

Twitter handle	Number of followers	Type of account	Geographic scope	Associated location
GovRicketts	17656	Government/ Personal	Regional (state)	Nebraska
Matt_Davison	43.9k	Personal	Regional	Lincoln, Nebraska
SarahFiliKETV	2426	Personal	N/A	N/A
DuffelBagDustin	338	Personal	N/A	N/A
Barbi_Twins	33.2k	Personal	N/A	Malibu, CA
NEMAtweets	6590	Government	Regional (state)	Nebraska
NENationalGuard	5980	Government	Regional (state)	Lincoln, NE

<u>Phase 2: Analysis of nodes from largest community</u>

Blondel et al.'s (2008) modularity class algorithm identified 496 distinct communities within the entire network. Out of these 496 communities, the largest community contained a total of 905 nodes, 11.2% of all nodes in the network. Figure 2.2 is a visualization of this community, where the colour intensity and size of each node correspond to the node's degree.

Figure 2.2: Visualization of largest community

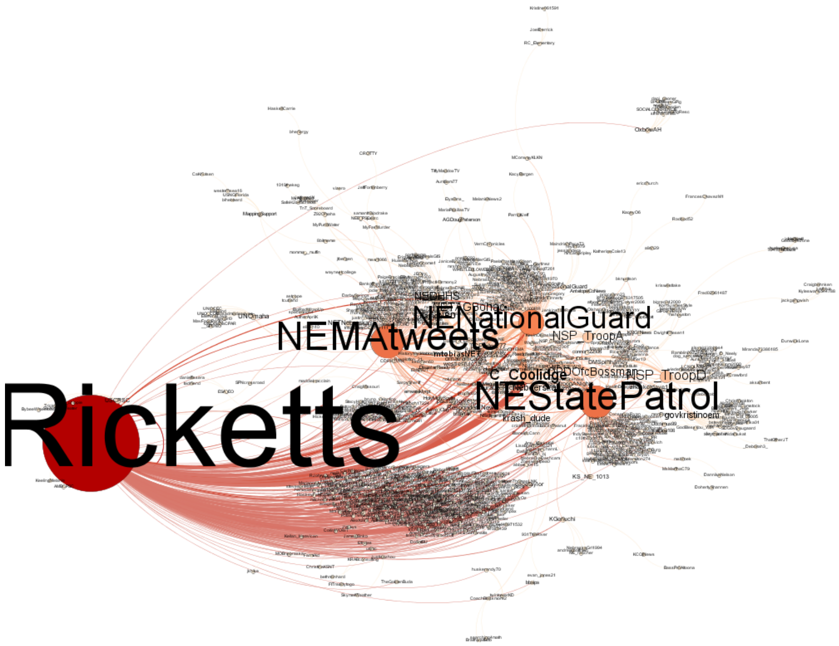


Table 2.4 shows the most central nodes in this community, according to degree centrality, betweenness centrality, and the PageRank algorithm. Accounts highlighted in blue are those that have top scores for all three centrality measures and accounts highlighted in green are those that have top scores for two of the three centrality measures.

Table 2.5 summarizes the characteristics of the nodes that appear in at least two out of the three top ten centrality rankings (those coloured in green and blue in Table 2.4). Each of the Twitter accounts listed in Table 2.3 may be understood as a significant source of information during the Nebraska flood event.

Table 2.4: Most central nodes in largest modularity class, according to three measures

	Degree	**Betweenness**	**PageRank**
1	GovRicketts	GovRicketts	GovRicketts
2	NEMAtweets	NETAGBohac	NEMAtweets
3	NEStatePatrol	NEMAtweets	NENationalGuard
4	NENationalGuard	NENationalGuard	NEStatePatrol
5	NSP_TroopD	NEStatePatrol	NETAGBohac
6	NETAGBohac	mtobiasNET	DouglasCountyNE
7	openmemories	NSP_TroopD	DCEMA_Nebraska
8	NSP_TroopA	femaregion7	govkristinoem
9	NEDHHS	NSP_TroopA	UNOmaha
10	C_Coolidge	fema	NSP_TroopA

Table 2.5: Characteristics of most central nodes in largest modularity class

Twitter handle	**Number of followers**	**Type of account**	**Geographic scope**	**Associated location**
GovRicketts	17656	Government/ Personal	Regional (state)	Nebraska
NEMAtweets	6663	Government	Regional (state)	Nebraska
NEStatePatrol	26.1k	Government	Regional (state)	N/A
NENationalGuard	5997	Government	Regional (state)	Lincoln, NE
NSP_TroopD	3369	Government	Local (counties)	North Platte, NE
NETAGBohac	535	Government/Personal	Regional (state)	N/A
NSP_TroopA	4170	Government	Local (counties)	Omaha, NE

Similar to the findings presented in Table 2.3, Table 2.5 shows that the Twitter account from Governor Ricketts has the highest scores across all three measures of node centrality. The accounts “NEMAtweets” and “NENationalGuard” are also included in both Table 2.3 and Table 2.5. Including the Governor’s account, there are a total of six Twitter accounts that appear in the top ten across all three measures of node centrality (as shown in Table 2.4). A single account appears in two of the top rankings.

The characteristics of the highly central nodes in this subset of the network suggest that the information (ie. Tweet content) being shared amongst this group of Twitter accounts is more formalized news about the Nebraska flood. Many of the accounts in Table 2.5 are official government accounts, rather than personal or commercial accounts. This finding may indicate that, within this cluster of tweets, information about the flood was communicated through official news sources. Some of the accounts identified in Table 2.5 communicate information at a local (county) level, suggesting that the information contained in the Tweets is potentially more specific to local impacts of the flooding. All accounts with associated locations are within Nebraska, further verifying that the Tweet content is relevant to the Nebraska flooding.

SECTION 3: Discussion

3.1: News sources during the Nebraska flood event

This analysis is an exploration into the social structure of Twitter interactions during an extreme weather event and the ways that information is communicated within a group of actors. Across both Phase 1 and Phase 2 of the analysis, the Twitter account associated with Governor Pete Ricketts of Nebraska was found to be the most influential source of information during the Nebraska flooding. The findings from Phase 1 suggest that many informal news sources (ie. personal Twitter accounts) may have been present during the Nebraska flood event. However, official news sources (such as those from the government), may have had a stronger presence within a subset of the network. Interestingly, no commercial news sources were identified in this analysis.

In future research efforts, this analysis could be extended to include a more rigorous and comprehensive characterization of the notable Twitter accounts. For example, perhaps future work could include a more rigorous characterization of the content associated with each account (ie. news source). Qualities such as topic of tweets, use of media, number of engagements, and commonly used terms could be analyzed to further investigate the characteristics of Twitter accounts that are found to be notable news sources.

3.2: Limitations

Timeliness of data collection

Important data may be missing from this dataset of Tweets as data collection began several days after the Nebraska floods had begun. In future efforts, it may be more appropriate to retroactively collect data, as the temporal bounds of the weather event under study may be better defined.

Selection of hashtags

The quality and relevance of the data is largely contingent on the appropriateness of the hashtags used to filter the incoming stream of Tweets. Prior to data collection, the manual search on Twitter for relevant hashtags was not an exhaustive or entirely systematic process. It is possible that some key hashtags were not detected, resulting in a dataset that may be missing some relevant data. Similarly, it is possible that some of the hashtags employed in data collection efforts were used in ways unrelated to the flood event. For example, a manual review of some of the Tweets included in the dataset indicates that the hashtags "#nebraskastrong" may also be frequently used in Tweets that refer to sports activities such as football. For future work, a more deliberate methodology for selecting hashtags prior to data collection should be developed. Additional strategies for filtering incoming Tweets, such as by geolocation, may also be explored.

Removal of noise

As an analysis of the text content contained in the Tweets in this dataset has not yet been conducted, it is not possible to verify whether or not the influential/central Twitter accounts identified in this network analysis are in fact Tweeting useful information or news relating to the Nebraska flood event. Due to the data collection approach, outlined in Section 1.1, all that can be confirmed is that all Tweets contain at least one out of a list of relevant hashtags. It is possible that a certain amount of noise is present in the dataset, perhaps including Tweets that offer condolences or general statements of personal sentiment regarding the flood. Future analysis should incorporate the findings from activities such as topic modeling to remove noise in the dataset due to Tweets that are irrelevant (ie. not news about the weather event in question).

Measuring Twitter interactions

Tracking interactions between Twitter accounts (retweets, mentions, and replies) is an imperfect approach for modeling the transmission of information on social media. There are likely many Twitter users who consume news from other Twitter accounts without leaving a traceable interaction in the form of a retweet, mention, or reply. Further analysis of Tweet content to identify linguistic similarities or shared use of key terms between known "news" accounts and other accounts could perhaps be used to trace transmission of information.

Assessing node centrality

The related body of literature does not offer a well-defined methodology for identifying central nodes in this research context, perhaps due to the variety of measures that can be used to assess node centrality. At a theoretical level, it is not intuitive which measure best corresponds to the context of information dissemination during extreme weather events. Further reading in this domain should be done to inform future work.

It is also important to note that the approach described in this report to identify central nodes does not result in a dichotomy between nodes that are central and nodes that are peripheral. The central nodes identified in this analysis are only considered to be central relative to other nodes in the network.

REFERENCES

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *science*, *323*(5916), 892-895.

Chatfield, A. T., & Brajawidagda, U. (2012). Twitter tsunami early warning network: a social network analysis of Twitter information flows. In *ACIS 2012: Location, location, location: Proceedings of the 23rd Australasian Conference on Information Systems 2012*(pp. 1-10). ACIS.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab.

Willis, A., Fisher, A., & Lvov, I. (2015). Mapping networks of influence: tracking Twitter conversations through time and space. *Participations: Journal of Audience & Reception Studies*, *12*(1), 494-530.

Yang, J., & Counts, S. (2010). Predicting the Speed, Scale, and Range of Information Diffusion in Twitter. *Icwsm*, *10*(2010), 355-358.