# On Quantifying Diffusion of Health Information on Twitter

Gokhan Bakal[1] and Ramakanth Kavuluru[2]

***Abstract*— With the increasing use of digital technologies, online social networks are emerging as major means of communication. Recently, social networks such as Facebook and Twitter are also being used by consumers, care providers (physicians, hospitals), and government agencies to share health related information. The asymmetric user network and the short message size have made Twitter particularly popular for propagating health related content on the Web. Besides tweeting on their own, users can choose to *retweet* particular tweets from other users (even if they do not follow them on Twitter.) Thus, a tweet can diffuse through the Twitter network via the follower-friend connections. In this paper, we report results of a pilot study we conducted to quantitatively assess how health related tweets diffuse in the directed follower-friend Twitter graph through the retweeting activity. Our effort includes (1). development of a retweet collection and Twitter retweet graph formation framework and (2). a preliminary analysis of retweet graphs and associated diffusion metrics for health tweets. Given the ambiguous nature (due to polysemy and sarcasm) of health relatedness of tweets collected with keyword based matches, our initial study is limited to $\approx$ 200 health related tweets (which were manually verified to be on health topics) each with at least 25 retweets. To our knowledge, this is first attempt to study health information diffusion on Twitter through retweet graph analysis.**

## I. Introduction

Over the last decade, online information sharing and consumption have become popular through social networks such as Facebook, Twitter, and Instagram and forums such as Reddit and Quora. According to a 2015 survey by Pew Research Internet Project [1], 76% of online US adults use a social networking site. Consumers are increasingly using online resources for gathering health related information. A 2013 survey [2] shows that 35% of US adults use the Internet to perform initial diagnosis of a medical condition that they or someone they know might have. Estimates based on Twitter based disease surveillance have been shown to align well with those obtained through conventional methods [3], [4] providing near real time insights and trends, specifically in early stages of epidemic outbreaks [5]. Past attempts also demonstrated Twitter's potential in improving health literacy [6], promoting fitness activity [7], tracking drug safety [8], and surveilling emerging tobacco products [9], [10]. Although these content based studies are excellent use-cases of microblogging websites for public health, a crucial but surprisingly unexplored phenomenon is that of diffusion of health information on Twitter. Our main goal here is to study properties of diffusion of health related tweets as observed in retweet graphs (also typically called *information cascades*).

As pointed out by Taxidou and Fischer [11], who provide one of the first analyses of robust reconstruction of retweet graphs[1], constructing retweet graphs in real time is a highly complex task owing to rate limits imposed by Twitter Inc. This involves two different non-trivial tasks that are affected by rate limits – first obtaining all retweets for a popular tweet up to a certain point in time and next getting social graph information (follower-friend connections) among retweeters to chart out potential paths using which the information has spread across the Twitterverse. Although other researchers have continued to analyze diffusion through retweet graphs in fields such as politics, crisis response, and sports [12]–[15], similar results are not available for health topics.

## II. (Re)tweet Collection Framework

Our goal is to design and implement a software framework that lets users identify popular tweets on specific topics, allows them build and visualize the corresponding retweet graphs, and compute certain measures of diffusion (more later) on those graphs. The overall design of the project is shown in Figure 1 with six numbered components identified in ovals in the figure whose brief details (due to space constraints) are as follows:

1) A user first specifies different sets of keywords (phrases) pertaining to specific topics of interest.
2) Twitter4J (`http://twitter4j.org`) interface to Twitter API is used to collect tweets that are stored in a MySQL database. Retweets in the stream are identified and the corresponding original tweets and their retweet counts are updated.
3) A process that continuously monitors the retweet counts and for tweets that have at least 25 retweets, uses API to get the corresponding new retweeter usernames periodically. Thus we impose a minimum threshold of 25 to determine popularity and this can be fine tuned depending on the domain.
4) After 7 days following a popular tweet, we stop the retweet collection and obtain all the follower-friend connections between the retweeters. We realize that there might be more retweets after the 7-day threshold but it has been shown that most retweets occur in the first one or two days and the associated counts follow an exponential decay rate [16].

[1]G. Bakal is with the Department of Computer Science, University of Kentucky, Lexington, KY, USA. `mgokhanbakal@uky.edu`

[2]R. Kavuluru is the *corresponding author* and is with the Division of Biomedical Informatics (Department of Internal Medicine) and the Department of Computer Science, University of Kentucky, Lexington, KY, USA. `rvkavu2@uky.edu`

[1]Our work was also conducted around the same time as part of the M.S. degree project of the first author

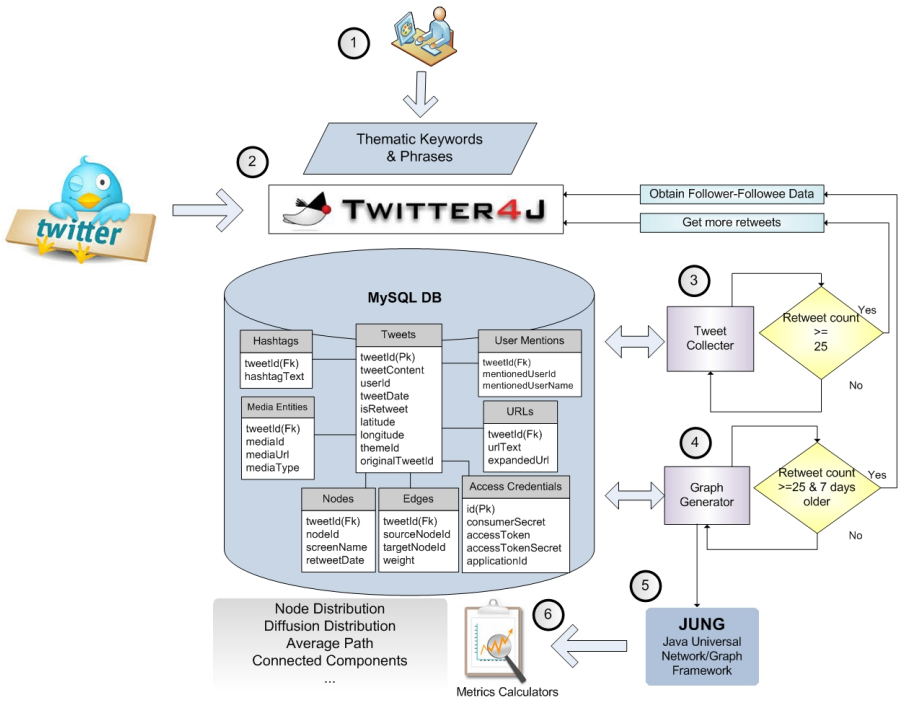


Fig. 1: Schematic of overall (re)tweet collection infrastructure

5) For any popular retweet, the corresponding retweeter connection subgraph (or more constrained versions of this subgraph) can be visualized using the JUNG open source graph analysis software [17].
6) In the end various metrics of interest can be computed on these retweet based graphs to measure information diffusion (more later).

For our current purposes we used a broad variety of health related themes as indicated by the key terms shown in the Appendix. Furthermore, since automatically identifying health related tweets simply using keywords is difficult (owing to homonymy, polysemy, sarcasm), we make sure that tweets also contain a URL in addition to having at least 25 retweets from component 3 in figure 1. This is because, in general, URL containing tweets were observed to contain more reliable health information as opposed to rumors and fake news. For specific details of the database tables and other subtler efficiency and software engineering related considerations, please see our comprehensive technical report (`https://goo.gl/x3Urz9`).

## III. Retweet Graph Generation

Once a retweet that is older than a week is identified (based on periodic scanning of the database), we use Twitter API to fetch the followers of each of retweet handles for each tweet. For these tweets, we thus have all the retweets (nodes table) and the corresponding follower lists. From the follower lists we can establish directed edges (edges table) among the set of nodes including the retweeter nodes and the root node. There are a couple of interesting ways we do this and we explain those graph construction methods here. Before that we explain our basic retweet graph construction.

### A. Basic Retweet Graph

The basic retweet graph $R_t = (V_t, E_t)$ for a tweet $t$ is built using the nodes $V_t = \{r(t)\} \cup \{u : u \text{ retweeted } t\}$, where $r(t)$ is the root node representing the original tweeter. Directed edges in $E_t$ are directly obtained based on follower-friend connections between nodes in $V_t$ based on certain constraints. For this basic retweet graph, $(u, v) \in E_t$ if and only if $u, v \in V_t$, $u$ follows $v$, and the retweet times satisfy the condition $\tau(u) \geq \tau(v)$, where $\tau(x)$ is the time at which $x$ retweeted $t$. This basic retweet graph represents all directed paths from the root to all retweeters. Since users can retweet any public tweet even if they are not following the original tweeter, this retweet graph may not be connected.

### B. Candidate Diffusion Graphs

While the basic retweet graph $R_t$ represents all paths through which a tweet could have reached a retweeter from the root, we also wanted to apply specific constraints based on time and distance that reduce the number of paths to at most a single path.

We define a time based candidate diffusion graph $D_t^\tau = (V_t^\tau, E_t^\tau)$ where $V_t^\tau = V_t$, the same as in the basic retweet graph. However, $E_t^\tau \subseteq E_t$ is constructed by removing some edges in $E_t$ based on the time of retweets. Specifically, retain only a single edge $(u, x)$ from a node $u$ by picking

$$x = \underset{y:\,(u,y)\in E_t}{\arg\max}\ \tau(y). \tag{1}$$

The intuitive basis for this is that a retweeter might choose to retweet because in their feed they will see the latest retweet first before they see retweets by other people they are following. Note that this approach generates trees unlike in

the basic graph and potentially more connected components than in the basic graph.

On the other hand, an argument can also be made that in a connected component, the earliest retweeter who $u$ is following might have encouraged $u$ to retweet it. This is because as $u$ encounters several retweets of the same tweet in her feed, she might pay more attention and attach more value as she notices the oldest retweets. That is, in equation 1, $\arg\min$ could also be used. Because of the way basic retweet graph $R_t$ is built, this naturally means the shortest path to a retweeter is chosen in each connected component form the corresponding root. We call this the distance based candidate diffusion graph $D_t^{\delta}$. All these retweet graphs are only 'candidates' because we cannot exactly determine what made the retweeters choose to retweet the original tweet although the candidates graphs throw light some plausible explanations. Readers are welcome to refer to the full technical report (`https://goo.gl/x3Urz9`) for example retweet graphs generated from our datasets.

### C. Diffusion Metrics

In this section, we discuss some of the metrics we think are appropriate for understanding retweet based diffusion on Twitter. Additional measures can be incorporated using the JUNG graph representation.

- **Node Distribution:** Node distribution is a sequence of values $P_j$ in $[0, 1]$ defined as

$$P_j = \frac{\text{\# nodes at level } j}{\text{\# total retweets}}, \tag{2}$$

  for the root component of diffusion graphs based on time or distance based constraints. Intuitively, this sequence gives insights into the penetration of a tweet in the follower-friend network.
- **Root Connected Component Contribution:** For each tweet $t$,

$$C_t = \frac{\text{\# nodes in the root component}}{\text{\# total retweets}},$$

  measures how many retweeters are connected to the root. We also count the numbers of connected components and also sizes of all components.
- **Average Number of Paths:** Since the basic retweet graph allows multiple paths of diffusion for a tweet from root to a retweeter, we wanted to measure the average number of paths to a retweeter in the root component for each tweet. The average number of paths for $R_t$

$$AP(R_t) = \frac{\sum_i (\text{\# paths to retweet node } i)}{\text{\# total retweets}} \tag{3}$$

## IV. Experiments and Analysis

Before we go into the details of some experiments we conduct, we discuss an important caveat here. Although our methods get most of the retweets, it is not always possible to get all retweets. This is indicated by the ratio of retweets we get and the actual retweet count reported by Twitter for each of tweets. In this analysis we only focus on tweets where this ratio at least 1. Note that the ratio can be greater than 1 because the actual retweet count may be larger than the retweets we get because some retweeters delete their retweets, which is not captured by our methods. When this ratio is less than 1, we would be missing some retweets and hence the number of connected components might not be accurate as the missed retweet could potentially make the graph more connected. Similarly, due to the delay between when the retweets are generated and when we get the follower-friend connections (the seven day threshold), some retweeters who were originally following some other nodes in $V_t$ might have chosen to unfollow. This could also render the graph less connected than it should have been at the time the retweet activity occurred. However, we assume that most of the connections remain intact at least in the first week after the original tweet.

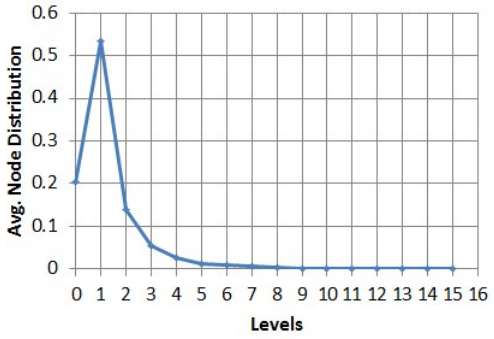


Fig. 2: Average node distribution in $D^{\tau}$ retweet graphs

Polysemy and other ambiguity issues plagued our keyword based tweet collection approach as expected (e.g., "great depression" vs "tropical depression" vs "chronic depression", all are hits when searching for "depression"; or "cancers dont talk about their problems" when searching for "cancer"). Jokes and sarcasm also contribute to the noise in keyword matches (e.g., "everybody has an addiction, mine is you" or "her smile could cure cancer"). The problems also persisted even after we required the presence of a URL. So although we had a total of nearly 50 million tweets on topics shown in the Appendix, we simply had to manually verify health relatedness of each tweet. We randomly selected two hundred popular tweets with URLs and manually determined that 179 are health related. Our analysis in the rest of this section is for this subset whose average retweet count was close to 100 with the maximum depth of the retweet graphs being 6.

The average node distribution at each level over all root components of all $D^{\tau}$ retweet graphs is shown in figure 2. On average, 53% of the retweets are from the followers of the root component (level 1). The root level (level 0) and the next two levels contribute to 87% of all retweets in the root component. We notice that 72% of the graphs satisfy $P_j \geq P_{j+1}$ for all $j$ (from equation 2), meaning three fourths of the time nodes at levels closer to the root are more in number than those away. There are few exceptions when $P_1 < P_2$ (7 cases) & $P_1 < P_2 < P_3$ (3 cases) which happened when nodes with many more followers than root follow the root and hence most diffusion occurs at deeper levels.

Next we study the proportion of nodes in the root connected component, number of connected components, and their distribution in $D^\tau$ retweet graphs. We notice that average proportion of retweets that belong to root component is 0.6. Thus, 40% of the time, users are searching for tweets and retweeting them even if they are not following the original tweeter. Only 8% have exactly one component and nearly 10% have as many components as there are retweets (so all singleton components). Average connected component size is 75. 21% of the graphs have a component larger than the root component. This shows that the original tweeters can gain greater influence by convincing the main nodes of these other components to follow and retweet their tweets.

Next we look at the average number of paths (equation 3) to a retweeter in the root component. The average of all $AP(R_t)$ over all $t$ is 37. The maximum AP value is 3475 for a graph which has only 26 retweets and is densely connected. 28% of the dataset has $AP$ value 1, which means there is exactly one path from the root to all retweeters in these cases. Clearly, for better diffusion, we would want $AP$ as high as possible. So having collaborators and other stakeholders of health organizations that seek to maximize diffusion of preventative health information make take strategic positions in the Twitter network to help achieve high APs.

## V. Conclusion

In this short paper, we described a framework to automatically identify popular tweets (based on keyword matches) and obtain the follower–friend connections involving the corresponding retweeters. Asynchronously, the framework uses the JUNG framework to visualize the retweet graphs. Since the graphs are meant to be modeling tweet diffusion, we derive a basic retweet graph based retweet times. Additional graphs that model diffusion based on time and distance are also created. Using the JUNG graph representation diffusion related analyses were conducted and the corresponding implications were discussed. Our framework and approach still have a couple of important limitations. First if we do not get the full set of retweets, which happens many times, we need a way to identify missing nodes that might make the retweet graphs more connected. One straightforward idea is to look at the followers of the author of popular tweet who are being followed by retweeters outside the main connected component. These immediate followers of the tweet author could be the missing retweeters that might offer a partial explanation for highly fragmented retweets graphs. Second, manual analysis is often needed to identify health related tweets since keyword based selection is bound to result in many false positives due to polysemy and homonymy. We are currently working on machine learning techniques that will aid in automatic health related tweet identification.

## References


[1] Pew Research Center. Social media usage: 2005–2015. http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/.

[2] Pew Research Internet Project. Health online 2013. http://www.pewinternet.org/2013/01/15/health-online-2013/.

[3] P. Velardi, G. Stilo, A. E. Tozzi, and F. Gesualdo, "Twitter mining for fine-grained syndromic surveillance," *Artificial Intelligence in Medicine*, vol. 61, no. 3, pp. 153–163, 2014.

[4] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing Twitter for public health." in *Proceedings of the Fifth AAAI Intl. Conference on Weblogs and Social Media (ICWSM)*, 2011, pp. 265–272.

[5] E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: Detecting influenza epidemics using Twitter," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11. Association for Computational Linguistics, 2011, pp. 1568–1576.

[6] H. Park, S. Rodgers, and J. Stemmle, "Analyzing health organizations' use of Twitter for promoting health literacy," *Journal of health communication*, vol. 18, no. 4, pp. 410–425, 2013.

[7] R. Teodoro and M. Naaman, "Fitter with Twitter: Understanding personal health and fitness activity in social media," in *Proceedings of the Seventh AAAI International Conference on Weblogs and Social Media (ICWSM)*, 2013, pp. 611–620.

[8] C. C. Freifeld, J. S. Brownstein, C. M. Menone, W. Bao, R. Filice, T. Kass-Hout, and N. Dasgupta, "Digital drug safety surveillance: Monitoring pharmaceutical products in Twitter," *Drug Safety*, vol. 37, no. 5, pp. 343–350, 2014.

[9] S. Han and R. Kavuluru, "Exploratory analysis of marketing and non-marketing e-cigarette themes on Twitter," in *Proc. of the 8th Intl Conference on Social Informatics*. Springer, 2016, pp. 307–322.

[10] R. Kavuluru and A. Sabbir, "Toward automated e-cigarette surveillance: Spotting e-cigarette proponents on Twitter," *Journal of biomedical informatics*, vol. 61, pp. 19–26, 2016.

[11] I. Taxidou and P. M. Fischer, "Online analysis of information diffusion in twitter," in *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 1313–1318.

[12] Z. Zhou, R. Bandari, J. Kong, H. Qian, and V. Roychowdhury, "Information resonance on Twitter: watching Iran," in *Proceedings of the 1st workshop on social media analytics*. ACM, 2010, pp. 123–131.

[13] C. Hui, Y. Tyshchuk, W. A. Wallace, M. Magdon-Ismail, and M. Goldberg, "Information cascades in social media in response to a crisis: a preliminary model and a case study," in *Proceedings of the 21st Intl. Conference on World Wide Web*. ACM, 2012, pp. 653–656.

[14] P. M. Fischer, I. Taxidou, B. Lutz, and M. Huber, "Distributed streaming reconstruction of information diffusion: poster," in *Proceedings of the 10th ACM International Conference on Distributed and Event-based Systems*. ACM, 2016, pp. 368–371.

[15] T. De Nies, I. Taxidou, A. Dimou, R. Verborgh, P. M. Fischer, E. Mannens, and R. Van de Walle, "Towards multi-level provenance reconstruction of information diffusion on social media," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 2015, pp. 1823–1826.

[16] T. Zaman, E. B. Fox, E. T. Bradlow *et al.*, "A bayesian approach for predicting the popularity of tweets," *The Annals of Applied Statistics*, vol. 8, no. 3, pp. 1583–1611, 2014.

[17] J. OMadadhain, D. Fisher, P. Smyth, S. White, and Y.-B. Boey, "Analysis and visualization of network data using jung," *Journal of Statistical Software*, vol. 10, no. 2, pp. 1–35, 2005.


## Appendix: Key Terms for Data Collection

We organized keywords used to collect tweets into the following groups representing different themes such as fitness, healthcare, chronic diseases, cancers, and tobacco products.

- fitness, nutrition, exercise, weightloss, workout, wellness, diet, gym, running, jogging, bodybuilding.
- healthcare, doctor(s), physician(s), healthinsurance, patient(s), pharmacy, therapy, hospital(s), medicine(s), disease(s), medication(s).
- obesity, diabetes, depression, alzheimer, addiction, mentalillness, arthritis, stroke, heartdisease, asthma, hypertension, epilepsy, copd, cvd.
- lung cancer, breast cancer, prostate cancer, colon cancer, colorectal cancer.
- smoking, tobacco, e[-]cigarette(s), snus, vape, vaping.