Information Diffusion in Online Social Networks

Adrien Guille

Information Diffusion in Online Social Networks

Adrien Guille
ERIC Lab, Lyon 2 University, France
5 av. Pierre Mendes France, 69676 Bron Cedex
adrien.guille@univ-lyon2.fr
Expected graduation date: Fall 2014
Advisors: Cécile Favre and Hakim Hacid, director: Djamel Zighed

ABSTRACT

Online social networks play a major role in the spread of information at very large scale and it becomes essential to provide means to analyze this phenomenon. Analyzing information diffusion proves to be a challenging task since the raw data produced by users of these networks are a flood of ideas, recommendations, opinions, *etc.* The aim of this PhD work is to help in the understanding of this phenomenon. So far, our contributions are the following: (i) a survey of developments in the field; (ii) *T-BaSIC*, a graph-based model for information diffusion prediction; (iii) *SONDY*, an open source platform that helps understanding social network users' interests and activity by providing emerging topics and events detection as well as network analysis functionalities.

Categories and Subject Descriptors

H.4 [**Information systems**]: Information systems application; I.6 [**Computing Methodologies**]: Simulation and modeling

Keywords

Online social networks, information diffusion

1. INTRODUCTION

Online social networks allow hundreds of millions of Internet users worldwide to produce and consume content. They provide access to a very vast source of information on an unprecedented scale. They also play a major role in the diffusion of information and have proven to be very powerful in many situations, like Facebook during the 2010 Arab spring [13] or Twitter during the 2008 U.S. presidential elections [14]. Still, the raw data produced by users of these networks are a flood of ideas, opinions, *etc.* and it becomes essential to provide means to analyze them.

As a computer scientist, I focus within the context of my PhD thesis on information diffusion in online social networks, and more specifically work or plan to work on the following issues under the guidance of my advisors: (i) *which pieces of information are popular and diffuse the most*, (ii) *how, why and through which paths information diffuses*, (iii) *which members of the network play important roles in the spreading process*, and finally, (iv) *how to build applications that exploit information diffusion in online social networks?*

Contributions. So far, we made the following contributions: (i) a survey of developments in the field, ranging from popular topic detection to information diffusion modeling, including influential spreaders identification; (ii) *T-BaSIC*, *i.e. Time-Based Asynchronous Independent Cascades*, a graph-based model for information diffusion prediction. Contrary to classical approaches where numerical parameters are fixed in advance, *T-BaSIC*'s parameters are functions depending on time, which permit a better modeling of what is observed in real-world social networks [11, 12]; (iii) *SONDY*, *i.e. SOcial Network DYnamics*, an open source platform that helps understanding social network users' interests and activity by providing emerging topics and events detection as well as network analysis functionalities. It also provides researchers an easy way to compare and evaluate recent techniques to mine social data, implement new algorithms and extend the application [10].

Basics of online social networks and information diffusion. An online social network results from the use of a dedicated web-service, often referred to as *social network site*, that allows its users to (i) create a profile page and publish messages, and (ii) explicitly connect to other users thus creating social relationships. *De facto*, an online social network can be described as a user-generated content system that permits its users to communicate and share information. An online social network is formally represented by a graph, where nodes are users and edges are relationships and can be either directed or not depending on how the social network site manages relationships. More precisely, it depends on whether it allows connecting in an unilateral (*e.g.* Twitter social model of *following*) or bilateral (*e.g.* Facebook social model of *friendship*) manner. Messages are published by the members of the network and constitute the main information vehicle in such services. A message is basically described by (i) a text, (ii) an author, (iii) a time-stamp and optionally, (iv) the set of people to whom the message is specifically targeted. Social network members publish messages to share or forward various kinds of information, such as product recommendations, political opinions, ideas, *etc.* Every piece of information can be transformed into a topic using one of the common formalisms detailed in Definition 1.

DEFINITION 1 (Topic). *A topic is defined as a co-*

SIGMOD'13 PhD Symposium, June 23, 2013, New York, NY, USA.

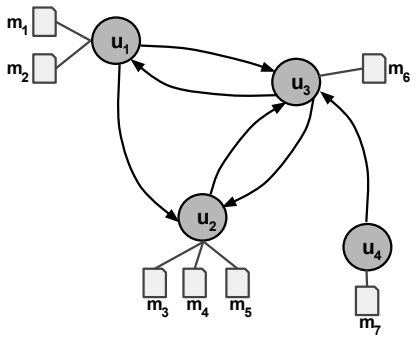

Figure 1: An example of social network enriched by the published content. Users are denoted u_i and messages m_j. An edge (u_x, u_y) means that user u_x is exposed to the messages published by user u_y.

herent set of semantically related terms that express a single argument. In practice, we find three interpretations of this definition: (i) a single term, (ii) a set of terms, (iii) a probability distribution over a set of terms.

Information diffusion is observed through the content generated by the network, namely a stream of messages. That stream can be viewed as a sequence of decisions (*i.e.* whether to react on a certain topic or not), with later people watching the actions of earlier people. Thus, individuals can be influenced by the actions taken by others. This effect is known as *social influence* [4], and is defined in Definition 2:

DEFINITION 2 (Social Influence). *A social phenomenon that individuals can undergo or exert, translating the fact that actions of a user can induce his connections to behave in a similar way.*

Figure 1 shows a directed social network comprised of four nodes with their related messages. This representation reveals that, for example, the user named "u_1" is exposed to the content produced by "u_2" and "u_3". It also indicates that none of the three other nodes are exposed to the information shared by "u_4". Figure 2 represents the stream of messages, *i.e.* a sequence of messages ordered according to the time axis, produced by these people.

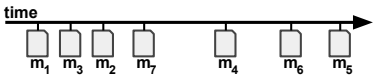

Figure 2: The stream of messages produced by the members of the network depicted on Figure 1.

Based on social influence, *herd behaviors* and *informational cascades* [15], respectively defined in Definition 3 and 4, have the potential to occur. In this context, some topics can become extremely popular, spread worldwide, and contribute to new trends.

DEFINITION 3 (Herd behavior). *A social behavior occurring when a sequence of individuals make an identical action, not necessarily ignoring their private information signals.*

DEFINITION 4 (Informational Cascade). *An informational cascade happens when people ignore their own private information signals and make decisions from inferences based on earlier people's action.*

Paper outline. The rest of this paper describes the contributions made so far and what remains to be accomplished. It is organized as follows: section 2 discusses the challenges this PhD work addresses and reviews related work. In section 3, we introduce *T-BaSIC*, a predictive model for information diffusion. Section 4 describes *SONDY*, an open source platform for social dynamics mining and analysis. Finally, we conclude and describe future work in section 5.

2. RELATED WORK

When studying information diffusion in online social networks, we must look at three key issues:

- Which pieces of information are popular and receive a lot of attention? This is in order to extract tables of contents to sum up on-going discussions, recommend popular topics to users, or predict future popular topics.
- How, why and through which paths information (i) has spread and (ii) will or would spread in the future? Knowing this is of outstanding interest to optimize online marketing campaigns, stop the spread of viruses, *etc.*
- Which members of these networks play important roles in the propagation process? Identifying the most influential spreaders in a network is critical for ensuring efficient diffusion of information.

In the following subsections, we review related work based on a survey we submitted to an international journal.

2.1 Detecting popular topics using diffusion

Leskovec *et al.* [17] show that the temporal dynamics of the most popular topics in social media are made up of a succession of rising and falling patterns of attention, in other words, successive focus and defocus on topics. As a result, information diffuses in a *bursty* way in online social networks. Many term-frequency-based methods have been developed to detect interesting topics that draw bursts of interest from a stream of topically diverse messages. Shamma *et al.* introduced *"peaky topics"* and *"persistent conversations"*, two normalized term frequency metrics. Lu *et al.* proposed to use the *"moving average convergence divergence"* (*MACD*) indicator to study topic trends. Still, this approach suffers from the fact that *MACD* is an inherently lagging indicator, since it relies on two exponential moving averages. Moreover, these methods define topics as single terms. That definition may not always be the most appropriate because of ambiguity and lack of context. AlSumait *et al.* [1] developed an *"on-line Latent Dirichlet Allocation"* (*OLDA*) that incrementally update its model. Thus it can track the evolution of richer topic definitions over time and detect emerging ones. However, *OLDA* is computationally expensive and can difficultly scale-up to the dimension of real-world scenarios. Recently, Takahashi *et al.* [19] proposed to detect bursty keywords by learning and modeling each user link creation behavior instead of analyzing term frequency.

2.2 Modeling information diffusion

It is possible to directly extract from the data *where* and *when* a piece of information propagated, but not *how* and *why* did it propagate. Several models have been proposed in order to capture the mechanics and dynamics underlying the diffusion process. We distinguish explanatory models – such as *NETINF* [8], *NETRATE* [7] or *INFOPATH* [9] – that aim at, given the complete observation of the diffusion of an information, retracing the implicit path taken by a piece of information, from predictive models. The latter aim at predicting how a specific diffusion process would unfold in a given network, from temporal and/or spatial points of view by learning from past diffusion traces. On the one hand, there are non graph-based methods, such as *SIS* [18] like methods which dynamics are described by differential equations, or the non-parametric *Linear Influence Model* (*LIM*) [20]. They are limited by the fact that they ignore the topology of the network and only forecast the evolution of the rate at which information globally diffuses. On the other hand, there are graph-based methods, like *Linear Threshold* (*LT*) [5] or *Independent Cascades* (*IC*) [6] based methods that are able to predict who will influence who. However they rely on the non-realistic assumption that diffusion unfolds in a synchronous manner along a discrete time-axis.

2.3 Identifying influential spreaders

Link analysis techniques have been used to identify influential spreaders, such as the *k-shell decomposition* [16], *log k-shell decomposition* [2] and *PageRank* [3]. Romero *et al.* suggested to enrich link analysis with nodal features like the rate of information forwarding and developed an extension of the well-known *HITS* algorithm, *Influence-Passivity*. Kempe *et al.* proposed to use *IC* and *LT* models to identify subsets of influential information spreaders. However, even if a lot of efforts have been done from the technical point of view, there is still a lot to do from the perspective of setting up a common and clear definition of the notion of influence.

2.4 Observations

Although several contributions exist towards information diffusion analysis, they are often limited by restricting assumptions. Some interesting axes of developments also remain ill explored. Moreover, researchers rarely provide implementations of their techniques, which makes it difficult to compare existing approaches and evaluate new ones.

In the rest of this paper, we describe our contributions regarding two issues: (i) information diffusion modeling and more particularly, how to predict that process; (ii) providing a tool to analyze information diffusion for researchers and end-users.

3. T-BASIC: PREDICTING THE DYNAMICS OF INFORMATION DIFFUSION IN ONLINE SOCIAL NETWORKS

State-of-the-art predictive models suffer from two issues: they are either (i) non graph-based and thus ignore important properties of online social networks or (ii) they don't handle time in an optimal way. We address these issues by developing a graph-based model of diffusion that fully integrates the temporal dimension. Contrary to classical approaches where numerical parameters are fixed in advance, *T-BaSIC*'s parameters are functions depending on time, which permit a better modeling of what is observed in online social networks (a fuller account of this work appears in [11, 12]).

3.1 Proposed method

Model formulation. *T-BaSIC* models the diffusion of information through a directed network $G = (U, E)$, where U is the set of all the nodes and $E(\subset U \times U)$ is the set of all the arcs. For each arc (u_x, u_y), there are two parameters: $f_{u_x,u_y}(t)$ that gives the probability that u_x transmits information to u_y at a time t of the day, with $0 < f_{u_x,u_y}(t) < 1$, and r_{u_x,u_y}, with $r_{u_x,u_y} > 0$. $f_{u_x,u_y}(t)$ is referred to as the *diffusion function* and r_{u_x,u_y} is referred to as the time-delay parameter. $f_{u_x,u_y}(t)$ is a function of nodes, edge and exchanged content features. As for the Independent Cascades (*IC*) model [6], the diffusion process starts from a given set of initially activated nodes S, but by cons, unfolds in continuous-time. Each node u_x that becomes activated at time t is given a single chance to activate each of its inactive neighbors u_y with probability $f_{u_x,u_y}(t)$. If the activation is successful, the distant node becomes active at time $t+r_{u_x,u_y}$. The stopping condition of the process is when no more activations are possible. Figure 3 illustrates this principle and shows the input and output of *T-BaSIC*.

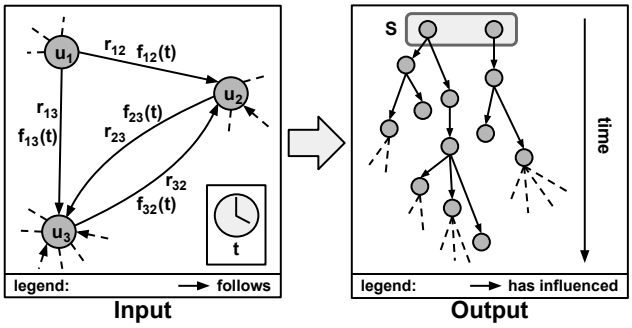

Figure 3: The T-BaSIC model predicts the diffusion process along a continuous time-axis based on the time-delay and diffusion function on each arc, starting from a set S of initially activated nodes.

Feature space. The model computes the probability that a node u_x transmits a piece of information i to node u_y at time t of the day. This probability is a function of nodes, edge and topic features belonging to the social, topical and temporal dimensions. These 13 interpretable features, which we describe below, are numerical values varying between 0 and 1 calculated on past information diffusion traces.

- *Social dimension features:* the rate at which each node publishes messages, $\mathrm{I}(u_x)$, $\mathrm{I}(u_y)$; a Jaccard coefficient of similarity between the two sets of nodes u_x and u_y interact with, $\mathrm{H}(u_x, u_y)$; the ratio of directed messages versus non-directed messages published by each node, $\mathrm{dTR}(u_x)$, $\mathrm{dTR}(u_y)$; the rate at which each node receives targeted messages, $\mathrm{mR}(u_x)$, $\mathrm{mR}(u_y)$;
- *Topic dimension features:* the interest of each user for the information, $\mathrm{hK}(u_x, i)$, $\mathrm{hK}(u_y, i)$;
- *Temporal dimension features:* the distribution of each user activity across the day, a non-parametric function stored as a vector, $\mathrm{A}(u_x, t)$, $\mathrm{A}(u_y, t)$;

Model parameter estimation. The diffusion probability, $f_{u_x,u_y}(t)$, is given by the following formula, where V is the related vector of features:

$$f_{u_x,u_y}(t) = P(\text{“diffusion”}|V) = \frac{1}{1+\exp(w_0 + \sum_{a=1}^{13} w_a V_a)}$$

The w_a coefficients are estimated using Bayesian logistic regression on data describing how information diffused in the network in the past.

3.2 Experiments

We evaluate the performance of *T-BaSIC* on a time series prediction problem with a Twitter dataset, for various topics and sub-networks. We compared the obtained results to the same baseline as the one used by Yang and Leskovec to evaluate *LIM*, namely the *one-time lag predictor* [20]. More specifically we studied the reduction over prediction error on two aspects, dynamics and volume. The evaluation demonstrated that *T-BaSIC* can more accurately predict the temporal dynamics of information diffusion than the *one-time lag predictor* (with an overall gain of precision of 32% on average) and *LIM*. However, it appears that *T-BaSIC* systematically under-estimates the volume. The main reason for that is surely the *“closed-world”* assumption underlying our modeling. *T-BaSIC* indeed only accounts for network influence and ignore external influence. Figure 4 shows an example of predicted time series compared to real time series for an information dealing with the possible release date of the next iPhone in two distinct networks.

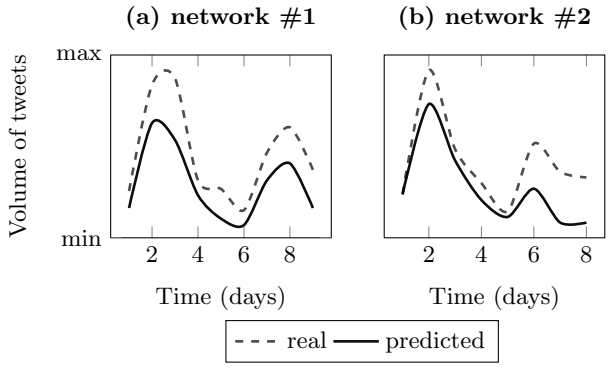

Figure 4: Comparison of real and predicted time-series for the topic {“iphone”,“release”} in two experimental Twitter networks.

4. SONDY: AN OPEN SOURCE PLATFORM FOR SOCIAL DYNAMICS MINING AND ANALYSIS

Although several contributions exist towards dynamics analysis in social data, most of them don't provide implementations of their techniques, and the few existing implementations are written in different languages and require different formatting and preparation of the input data, making it nearly impossible to compare the various approaches. Besides the difficulties around proposing new techniques for topic detection, these tasks necessitate generally a heavy pre-processing step which is performed manually.

Proposed platform. We propose *SONDY*, a tool that tackles the following two issues: (i) how to assist researchers and end-users in pre-processing the data, detecting topics and their trends, analyzing the corresponding networks (*i.e.* active authors for the considered topic(s)), and (ii) how to make it effortless to integrate, compare, and eventually combine different approaches to mine such data. *SONDY* is an open source platform integrating optimized implementations of some topics detection and graph mining algorithms in the same platform. It also provides researchers an easy way to compare and evaluate recent techniques to mine social data, implement new algorithms and extend the application (a fuller account of this work appears in [10]).

Platform Design. The application is written in JAVA and relies on four services to address the mentioned issues. Figure 5 shows numbered screen captures that illustrate the user interface of these services.

1. *Data manipulation service* (see Figure 5.1): for importing and preparing the data in order to optimize their exploitation and processing. This component includes stop-words removal, content stemming, message stream discretization, and message stream resizing.

2. *Topic detection and exploration service* (see Figure 5.2): for identifying and temporally locating trending topics and events. It encapsulates a set of algorithms for trends detection (such as *peaky topics, persistent conversations* and *MACD*[1]) combined with results visualization under several forms with different settings.

3. *Network analysis and visualization service* (see Figure 5.3): to observe the social network structure and find, *e.g.* influential nodes or communities. It encapsulates a set of algorithm for graph coloring (namely *k-shell decomposition* and *PageRank*[2]) and interactive visualizations, making it possible for users to actively interact with the system.

4. *Extension manager*: for importing new algorithms to be used by the the topic detection or network analysis services. This is done by importing JAR files.

5. CONCLUSION AND FUTURE WORK

In this paper, I summarize my up-to-date PhD thesis work. I first describe the challenges of information diffusion in online social networks study and present related work. Then I introduce the contributions made in this work w.r.t the state-of-the-art, namely a predictive model for information diffusion and an open source platform for social dynamics mining that will be used to implement future development of this thesis.

My current work focuses on summarizing information in online social networks. To this end, I am studying how to extract intelligible “table of contents” (*i.e.* not simply define a topic as a single term) while preserving computational efficiency (so the solution can scale-up to real-world scenarios), using mainly time series analysis techniques. The underlying assumption is that the “mentioning” frequency is a better indicator of the popularity of a topic than its global frequency. Mentions are links that are dynamically created

[1]See Section 2.1.
[2]See Section 2.3.

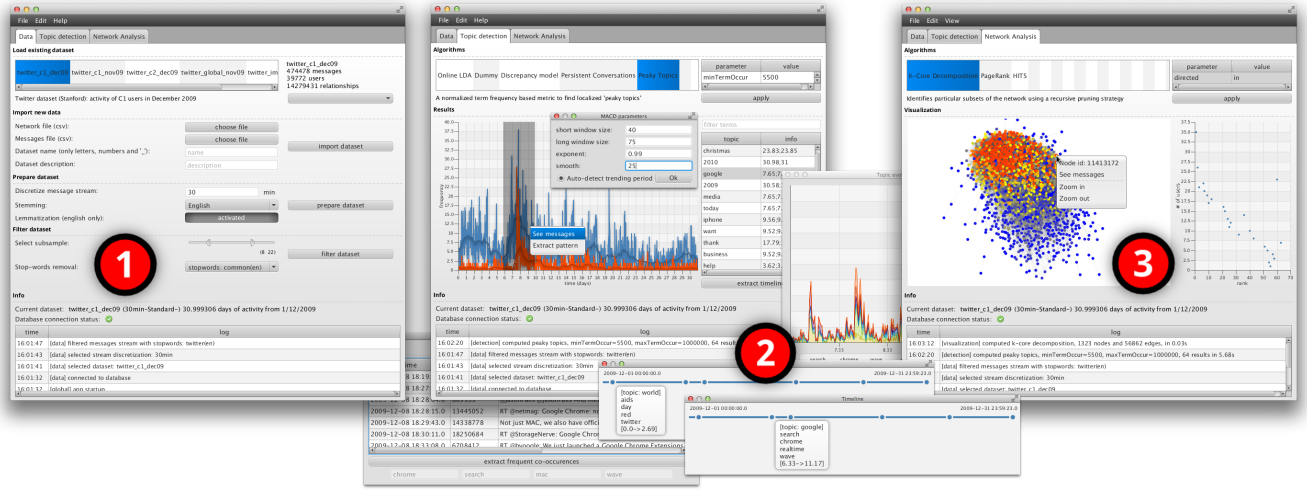

Figure 5: Illustration of the different services offered by *SONDY*, from left to right: (1) the data manipulation service, (2) the topics and trends exploration service and (3) the network analysis service.

between users, in the case where the author of a message wants to target it to one or more specific users, when replying to someone or "re-tweeting" a message for instance[3].

6. REFERENCES

[1] L. AlSumait, D. Barbará, and C. Domeniconi. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *ICDM '08*, pages 3–12, 2008.

[2] P. Brown and J. Feng. Measuring user influence on twitter using modified k-shell decomposition. In *ICWSM '11*, pages 18–23, 2011.

[3] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *MDMKDD '10*, pages 5–14, 2010.

[4] E. David and K. Jon. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. 2010.

[5] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the twitterers - predicting information cascades in microblogs. In *WOSN '10*, pages 3–11, 2010.

[6] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 2001.

[7] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal dynamics of diffusion networks. In *ICML '11*, pages 561–568, 2011.

[8] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *KDD '10*, pages 1019–1028, 2010.

[9] M. Gomez-Rodriguez, J. Leskovec, and B. Schökopf. Structure and dynamics of information pathways in online media. In *WSDM '13*, pages 23–32, 2013.

[10] A. Guille, C. Favre, H. Hacid, and D. Zighed. Sondy: An open source platform for social dynamics mining and analysis. In *SIGMOD '13*, 2013.

[11] A. Guille and H. Hacid. A predictive model for the temporal dynamics of information diffusion in online social networks. In *WWW '12 companion*, pages 1145–1152, 2012.

[12] A. Guille, H. Hacid, and C. Favre. Predicting the temporal dynamics of information diffusion in social networks. *ERIC Lab Report*, RI-ERIC-13/001, 2013.

[13] P. N. Howard and A. Duffy. Opening closed regimes, what was the role of social media during the arab spring? *Project on Information Technology and Political Islam*, pages 1–30, 2011.

[14] A. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3):248–260, 2009.

[15] S. Kariv et al. Distinguishing informational cascades from herd behavior in the laboratory. *The American Economic Review*, 94(3):484–498, 2004.

[16] M. Kitsak, L. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. Stanley, and H. Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, Aug 2010.

[17] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD '09*, pages 497–506, 2009.

[18] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *SDM '07*, pages 551–556, 2007.

[19] T. Takahashi, R. Tomioka, and K. Yamanishi. Discovering emerging topics in social streams via link anomaly detection. In *ICDM '11*, pages 1230–1235, 2011.

[20] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *ICDM '10*, pages 599–608, 2010.

[3] `https://support.twitter.com/articles/14023-what-are-replies-and-mentions`