Online Analysis of Information Diffusion in Twitter

Io Taxidou, Peter M. Fischer
University of Freiburg, Germany
{taxidou,peter.fischer} @informatik.uni-freiburg.de

ABSTRACT

The advent of social media has facilitated the study of information diffusion, user interaction and user influence over social networks. The research on analyzing information spreading focuses mostly on modeling, while analyses of real-life data have been limited to small, carefully cleaned datasets that are analyzed in an offline fashion. In this paper, we present an approach for online analysis of information diffusion in Twitter. We reconstruct so-called *information cascades* that model how information is being propagated from user to user from the stream of messages and the social graph. The results show that such an inference is feasible even on noisy, large-scale, rapidly produced data. We provide insights into the impact of incomplete data and the effect of different influence models on the cascades. The observed cascades show a significant amount of variety in scale and structure.

1. INTRODUCTION

Social media such as online social networks (Facebook), micromessaging services (Twitter) or sharing sites (Instagram) provide the space in which a significant part of social interactions takes place. Many real-life situations like elections are reflected by social media and in turn social media shapes them by forming opinions or strengthening trends. In addition to providing a large audience, social media has changed the speed of interaction: Information spreads within minutes or hours, triggering equally fast reactions. This often overwhelms all participants: Companies as well as politicians are struggling to keep up with the onslaught of (mostly negative) reactions which come suddenly and with high speed.

As a result, monitoring social media in real time has attracted a lot of interest both by academia and industry, with a strong focus on sentiment analysis [15] and trend detection [2, 14]. More thorough analysis like an in-depth understanding of how information is being spread or what are the particular user roles, has been so far performed in an offline fashion, typically targeting only academic research. Yet, in many situations, such deep understanding is needed; companies, celebrities or politicians need to react quickly to a massive amount of opinions, determining who is reacting to certain information and who is influencing others. Similar analysis may be useful for online journalism, helping to detect events, assessing their source, predicting virality as well as determining the impact of own publications. Furthermore, online marketing will benefit from such real-time analysis to rate the effectiveness of its methods and, if necessary, to adapt the means to reach the appropriate users.

An important area of analysis - both on its own and as an underpinning for more complex analysis is the study of information diffusion, i.e. tracing, understanding and predicting how a piece of information is spreading. In this paper, we present methods and results of how information diffusion can be studied in real-time, using retweets on Twitter as a starting point. We tackle the problem of determining *influence paths* that express the relationship of "who was influenced by whom". The set of influence paths form a social graph, that share a common root (a single user who first seeded a tweet) is referred as "*information cascade*" in the literature [13]. Nodes of the cascade represent nodes (users) of the social network that got "influenced" by the root or another user. Edges of the cascade represent edges of the social graph over which influence actually spread. An "*influencer*" in the case of Twitter is the so called "friend" that exposes information to his/her followers and exerts influence on them in such a way that they forward this piece of information.

Our online method relies on an algorithm and supporting system to infer possible influence paths from the stream of messages (tweets) and the underlying social graph (follower and friend network). To our knowledge, no work exists on reconstructing information cascades and inferring influence paths online while investigating the impact of incomplete and not cleaned datasets on such evaluations. Such incomplete datasets derive from API limitations or lack of explicitly observable user influence. Our method can be used as a general model of inferring influence paths, not only restricted to retweets, but also of any kind of information that propagates over a social network, e.g. URLs or hashtags.

In detail, we provide insights in the follows areas:

- Social connections as carriers of information: Is information propagated mostly over explicit links (like friends or followers) in social media or do other means play an important role? If the latter is the case, tracing and attributing influence becomes challenging. The results show that a large amount of influence can indeed be attributed to explicit social links.
- Feasibility and quality of inference: When working with online datasets, we encounter problems such as missing messages and missing social graph information. We show that we can reconstruct such cascades and the results are meaningful under the constraints of online analysis.
- Properties of information spreading: We provide evaluation of our datasets with the following insights; how influential is

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2580050.

the root user? Cascades tend to be wide or deep? To what extend users are exposed to multiple influencers and what are the effects of various influence models?

The remainder of the paper is structured as followed: Section 2 provides more background information on relevant existing research. In turn, Section 3 describes our model and algorithm. Our dataset is explained in Section 4, while the results of our evaluation are presented in Section 5. The paper concludes in Section 6.

2. RELATED WORK

Information diffusion and information cascades have been studied in the past in an offline way with relatively small datasets. A summary of models and methods of information diffusion is described in [9]. Two baseline approaches presented there are the Independent Cascades (IC) and the Linear Threshold (LT) model. The IC model [7] includes a diffusion probability that is associated with each edge while the LT model [8] defines an influence degree on each edge and an influence threshold for each node. The statistical, structural and content aspects of information cascades have been studied in [16, 12, 10]. In [16] authors investigated the size, shape and decay factors of cascades; the biggest cascade in the evaluations dataset contained 1K messages. In [12] shape and temporal analysis of retweet cascades were analyzed with the biggest retweet cascade containing 4K messages. The authors of [10] investigated human interactions on a crisis constructing the corresponding cascades, using a tiny dataset containing 168 retweets.

Relevant to our research is the work by Cogan et al. [5] that studied user interactions on Twitter, designing an algorithm to reconstruct the conversational graphs (mentions, retweets, replies). Their dataset contained 33K retweets while the largest retweet cascade had a size of 170 retweets. An offline straightforward Map-Reduce algorithm for reconstruction of retweet cascades is described in a Stanford class project [6].

Evaluating information diffusion with missing data has seen some interest. In [11] the effects of missing data in social networks are studied, by building a model to estimate the overall properties of information cascades given only a sample of nodes and edges. [17] aims to infer missing nodes by incorporating temporal information, but the cost of such inference is quite significant, depending on the size of the entire social graph. However, we tackle a different problem, proving that even with limited data it is possible to reconstruct cascades.

As far as real time analysis on social media is concerned, research is restricted to the domain of trend detection and event identification [2, 14]. These methods focus mainly on topic extraction and not on user interactions, as a result they are not directly comparable to our real-time approach for studying user interactions.

Overall, the existing methods and models for information cascades are implemented offline in relatively small datasets. Moreover, these models are restricted to specific conventions for reconstruction (e.g. '@' for retweet) while our algorithm can be extended to any information that propagates over a social graph. To the best of our knowledge, no work exists on reconstructing information cascades and inferring influence paths on real-time while investigating the impact of missing information.

3. MODELS AND ALGORITHMS

In order to track information diffusion on real-time, we need to extract information cascades out of the message stream. A cascade is formed when users forward the same original message from a user that we call the *root user*. The exact influence path, that shows how forwarding occurs, is not available from the message stream. Under the assumption that the social graph connections (e.g. followers and friends) serve as means of information diffusion and influence, we can derive these influence paths from these social connections among users.

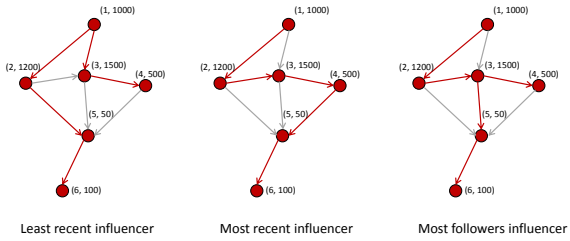


Figure 1: Impact of Influence Models

A core aspect of modeling information diffusion is the assignment of influence: users might be exposed and influenced by a piece of information by in multiple users, hence forming multiple influence paths [3]. When a message arrives that is a retweet, every friend that has (re)tweeted at an earlier point in time has to be considered as a potential influencer, if no constraints are made on the influence model. Specific influence models, however, may include only a subset of these influencers, as in reality users are not influenced equally by all their friends that forwarded the same message.

In order to investigate the mechanism of influence and avoid exhaustive search to all follower data that drives complexity, we employ different influence assignment models. We have considered the following models for influence assignment [3, 4]:

- *Least recent influencer*: Users are influenced by the first exposure even if they do not act immediately.
- *Most recent influencer*: Users are influenced by the last exposure.
- *Most followered influencer*: Users with the most followers tend to be more popular and it is assumed that they can trigger more retweets.
- *Most retweeted influencer*: Users whose messages are forwarded the most, emit interesting content and are considered to be authorities, thus exerting more influence on others.

We present an example in Figure 1 in order to better understand the aforementioned models and their impact on influence paths. Nodes represent users who propagate the same tweet while the arrow direction indicates the relationship *"is followed by"*. We apply three of the aforementioned influence models and construct the cascade accordingly. For each user there are two values attached: the first one indicates the temporal order of sending retweets and the second the number of followers. The influence paths in each case are highlighted in red. Note here that influence paths in red form a subset of the influence paths in the cascade, since these models trim the redundant influence paths. We can observe that each model (least recent/most recent/most followers influencer) leads to different structural and conceptual results thus exerting a big impact on the paths that information/influence flows.

In order to provide a clear understanding on the interaction of social network connections and messages streams with information cascades, we provide a lightweight formalization: The social graph $SG = (V, F)$ is a directed graph of follower/friend relationships, showing for each user (node) from V who follows this user (F). For simplicity we assume that during the reconstruction this social graph remains static. The message stream is expressed as sequence of messages M^* in temporal order. Each message M contains

several attributes, out of which we just list the ones most relevant for our work: (1) timestamp t (2) user $v \in V$ (3) information item identifier i, e.g. a retweet ID or a hashtag. We say that two messages m_1 and m_2 belong to the same cascade iff $m_1.i = m_2.i$. This model is flexible enough to express many kind of information diffusion, not just retweets.

Based on foundations, we define a cascade graph $C(U, E)$ with $U \subseteq V$ as directed graph of influence paths among users. This graph is a subset of SG annotated with (at least) the influence time on the edges. C contains only nodes of those users who actually (re)tweeted, but not those that were exposed to the information, but did not react. This way, we can use a smaller graph that focuses on the influence paths. If needed, a full reach computation can easily be achieved by incorporating these passive users that are followers of (re)tweeters. Among the users in this cascade that tweeted, we designate $u_r \in U$ as the "root", i.e. the user who initially distributes the information item.

The influence paths (edges) need to fulfill the following condition: an edge $(u_i, u_m, t) \in E, u_i, u_m \in U$ may only exist if $\exists m \in M^* : m.u_i = u_m \wedge (u1, u2) \in F \wedge (u_i = u_r \vee (\exists n \in M^* : n.u = u1 \wedge n.t < m.t))$. In other words, a user u_m who spreads information using a message m is possibly influenced by an user u_i if there is a social network connection from u_i to u_m and u_i is either the root or was exposed to this information by a message n which happened before m.

We design the baseline version of our reconstruction algorithm to exhaustively search theses edges in E for all messages in M^*, regardless of the used influence model. As our goal is to perform this reconstruction in an online fashion, these edges shall be added to C in an incremental manner whenever a message arrives. Our algorithm therefore checks at every arrival of a message m if a SG connection from $m.u$ to the user of any message $n \in M^*$ which arrived before m exists. If such a connection holds, it is a possible influence path and will be added to C. This leads to $\mathcal{O}(|M^*|^2)$ cost for the reconstruction of the entire cascade, if we assume a constant cost for checking the existence of an edge in SG. For specific influence models, we can utilize a refined version of this algorithm: only a limited amount of influence paths needs to be determined, so possibly only a few of the messages in M^* need to be checked. This may lead to lower average cost, but the specific reconstruction cost will depend on the influence model, the properties of the cascade and the representation of M^* .

When reconstructing information cascades from real data, we encounter either missing messages (nodes) or social graph connections (edges). In turn, this means that we have missing user nodes in C as well as missing influence edges, as these are being derived by the algorithm above. Furthermore, not all influence paths will actually be over explicit social network links. Instead, external influences or overviews on trending topics provide connections that are not captured by our approach. When the algorithm encounters a message that cannot be assigned to a previous influencer, this message becomes the root of a new fragment. As a result, we will not generate a single graph, but multiple fragments that are not connected to the main graph.

In order to implement evaluation in disconnected information cascades and compute metrics on them there are three ways to deal with the problem of disconnected fragments: (1) Considering only the large connected component or root component. (2) Evaluating the entire forest of all fragments in evaluation, but not joining them. (3) Inferring connections between fragments (nodes). In this case, we need to infer influence edges, which is discussed in the literature [11, 17], but no scalable methods exist. For our evaluations, we generally chose the second approach. For metrics that require a connected component (e.g. paths) we consider the largest connected component. In the future, we will consider and implement more elaborate models to connect cascade fragments.

For assessing the connectivity of information cascades, we introduced two metrics. Let $C = (U, E)$ be an information cascade graph with U being the nodes and E the edges, u_r the root and M^* a sequence of messages. To evaluate the connectivity of a diffusion graph the two formulas Connectivity-Rate (CR) and Root-Fragment-Rate (RFR) have been used.

$$\text{CR} = \frac{|\{u|(u', u) \in E \ \vee \ (u, u') \in E\}|}{|U|} \tag{1}$$

$$\text{RFR} = \frac{|\{u_j \in U|\text{iff exists a path } u_r, .., u_j \text{ in C}\}|}{|U|} \tag{2}$$

The Connectivity-Rate assesses whether there is a connection between two users (nodes) in the cascade. It returns the percentage of users that have at least one connection, and are thus influenced by another user. The Root-Fragment-Rate assess whether there is a path to the root user from every other user. It returns the percentage of nodes that are connected with the root directly or via an influence path over multiple users. CR provides a very basic and loose indicator, whereas RFR utilizes a very strict notion. Taken together, they provide sensible bounds for many more advanced metrics.

We implemented this model and algorithms on top of Storm [1], a scalable, distributed data stream processing platform which provides the necessary low-level primitives for distributed stream processing. Since we need to reconstruct with low latency and high efficiency, there is the requirement to store snapshots of the graph in distributed, main-memory storage components that support the required access patterns. To exploit locality and keep communication cost low, the social graph snapshots should be distributed to the expected computation distribution. Partitioning the social graph snapshots accordingly and investigating suitable systems to store the graphs are work in progress.

4. DATASET

4.1 Retrieval Approach

Performing an online analysis of information diffusion requires access to the relevant messages while they occur as well as an up-to-date instance of the social graph. For both goals, we need to overcome a number of challenges, requiring particular retrieval strategies. Among the popular online social media services, Twitter is the only one that provides an API to access messages and social graph information on the fly, but this API bears significant restrictions.

Messages.

For messages, Twitters' Streaming API[1] grants access to a subset of the current stream of messages. This subset can be defined on the basis of user names, keywords (including hashtags) and geocoordinates. There are, however, two kinds of restrictions on this API: On the one hand, the number of user names, keywords and coordinates that can be followed by an account are limited (currently to 5000 each). On the other hand, the number of messages per time produced by such a subscription must not exceed 1% of the total number of messages processed by Twitter at the same time. In cases of heavy traffic - such as a very popular topic at a certain instance - this threshold is exceeded, so we are missing messages

[1]https://dev.twitter.com/docs/streaming-apis

(and retweets). Furthermore, Twitter provides only limited means to retrieve messages after their occurrence.

These limitations have another consequence: we cannot observe all possible retweet cascades, but need to settle for specific subsets before we start to record. Generally speaking, we would need to perform some kind of event or virality detection on the fly in order to determine this subset, which is a research problem on its own. For the time being, we settled for two simple, but still promising approaches to achieve this goal: If we are aware of events that are likely to generate a considerable amount of tweets and retweets (such as Olympics 2012 or US elections 2012), we use specific keywords to track cascades referring to such events. This approach bears the drawback that we can request only messages of events known in advance. To overcome this problem and catch also emergent or unpredictable events, we observe the Twitter "sample" stream, containing a small randomly sampled subset of the full message stream. We detect relevant cascades that demonstrate a bursty behaviour in their beginning without knowing the specific topic of them. The beginning of the cascade is then immediately fetched using the Twitter REST API.

Social Graph.

For the social graph, Twitter offers methods to retrieve connections for every user, both the list of users who follow this user (followers) and the list of users this particular user follows (friends). Even compared to the limits on message subscriptions, the limits on the social graph are very strict: at most 60 users or 300K follower entries (whatever is smaller) can be retrieved per hour and account. Since we need to deal with high message rates in cascades, on-demand retrieval of current social network information during the reconstruction is not feasible. Instead, we have to retrieve the social graph over time, cache it and refresh it in order to reflect the graph evolution due to following and unfollowing of users over time. Given the sheer size of the social graph (100s of millions of users with their connections), we crawl the social by fetching information of those users that are active in retweets, with an emphasis on those users that are retweeted most and/or have the most followers, in order to capture possible popular users that exert influence on others [4]. When necessary, we can augment this collection by explicit requests on specific users. Since retrieving follower and friend information would provide redundant information, we chose to retrieve only the follower information. This is motivated by the fact that followers information provides a better expression of influence and gives a quick way to retrieve all connection information for the starter of a cascade.

4.2 Properties

Since our focus is to study realtime influence computation thoroughly in a reproducible manner, we still had to record a certain amount of data for evaluations that could be replayed. As a starting point (which we present here), we settled for a dataset that was recorded from August 3rd to September 24th 2012, covering most of the Olympics and the Paralympics 2012. Our analysis of other datasets is currently ongoing. We used the Twitter streaming API to subscribe to the filter terms "Olympics" and "London2012". In total the data set contains almost 11 million tweets, in particular 1.1 million separate retweet cascades - both values are significantly larger than any of datasets studied in the literature [16, 12, 10].We performed an initial analysis to understand some of the overall properties of this dataset, encountering a skewed distribution: the largest cascade has more than 60000 retweets, around 150 have more than 1000 retweets, approximately 5000 cascades have more 100 retweets and around 45000 cascades contain 10 or more retweets. Twitter includes a *retweet_count* field in every retweeted tweet, so we could compare the number of recorded retweets with the number of reported retweets for every cascade. For most of the cascades we recorded, these numbers showed only minor differences (around 15% on average). For 50% of the cascades we get 90% completeness or more, while for only 15% of the cascades we get completeness less than 80%. That means that our recording policy of tweets through the Twitter API works well. The only major exception happened during the “peak hours” of the Olympics, where the aggregate number of tweets from this subscription exceeded the 1% rate limit of Twitter, and matching tweets were dropped from the system. Our analysis also showed that messages were received in temporal order, so that we can process the message straight away without buffering and sorting.

In order to ensure a good coverage of the social graph, we ensured that the follower information of all the 1.2M users present in these cascades had been retrieved. There were two very distinct subsets of users present: For around 300K users, no follower information was accessible since these users have been blocked by Twitter or made their accounts private. For the remaining users, we fetched their followers, while the number of followers reported in the retweet message at the time of the recording have a close correlation. Since we fetched the followers after the recording of messages the number of users retrieved is slightly greater. This means that users follow more often than unfollow.

5. EVALUATION

In this section we evaluate our algorithm and models for reconstructing information cascades. The focus of this analysis is on data quality, feasibility and cascade properties, determining how interesting such analyses are and to which extend they yield useful results. For the real-time reconstruction performance, we present some initial insights: From an execution speed point of view, even a non-optimized implementation of the *complete* influence model finishes the reconstruction of large cascades in a few seconds (e.g. around 4 second for a cascade with 9000 users), if the social graph information is available in main memory on the same machine. Given the sheer size of the social graph, storing it in a single machine is not a very workable solution, so we are currently working on suitable models on how to compress the social graph and distribute it together with the computation.

The results we present on the cascade data cover four aspects, using the complete model for all but the last aspect: First, we confirm our assumption that social links are carriers of information. As a second step, we show how the quality of input data affects the reconstruction influence paths. Then, on top of reconstructed cascades we perform a preliminary analysis and compare it with previous studies on Twitter. As a fourth step, we investigate how different influence models described in Section 3 exert an influence on reconstruction of cascades.

5.1 Assignment of Influence

First, we evaluate our assumption that information flows through social links. We use a subset of our dataset which contains cascades with more than 100 messages (we call it ”full dataset”), due to the fact that the impact of incomplete data in small cascades has more unpredictable results in reconstruction rates. For testing reconstruction rates we used (1) a dataset containing more than 100 full dataset and (2) a cleaned subset of it. The cleaned dataset contains cascades with more than 90% of their messages acquired and having available more than 80% of follower lists. For the cleaned dataset, we get median connectivity rate 85% and root fragment rate 80%. When we extend our evaluations to the full

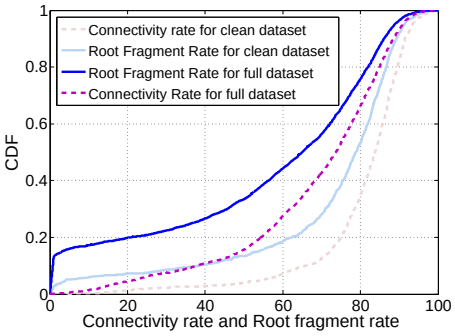


Figure 2: Reconstruction rates

dataset (that is dirty and incomplete) these rates drop. However, we show that it is possible under data limitations to reconstruct cascades and to obtain meaningful and decent results. For 20% of the cascades we get more than 80% connectivity rate and 70% root fragment rate (Figure 2). In ideal cases of message completeness 99% and follower lists 95% we get CR=93% and RFR=90%. As a result, we can conclude that social links are indeed the predominant carriers of information. However, there are still 10% messages that cannot be assigned using social graph information. That means, either the user has no social connections available (deleted or private account), or the user forwarded a message without having a direct link to any of the previous (re)tweeters (forwarded it from the public Timeline where messages of non followers are depicted).

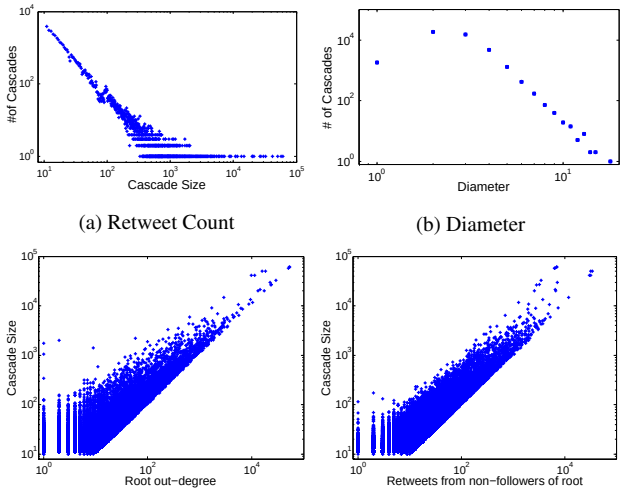
5.2 Impact of incomplete data

Next, we investigate explicitly the impact of different incompleteness parameters on the connectivity rate. Since we target online analysis, either messages are missing or social graph information might be absent or outdated. We take two cascades of size 1000, one star and one with a complex structure, with very good connectivity rates in the presence of full social network data and messages. In order to investigate the impact of incomplete data we removed gradually (1) follower lists, (2) messages. Due to space limitations, we are presenting only these two representative cascades, since results on other cascades are very similar.

For case (1) shown in the upper part of Table 1, we gradually removed follower lists apart from the root's, keeping the ones with the greater number of followers. Star cascades are expected to undergo lower degradation since most of the users (retweeters) are connected with the root. We can observe that by degrading the follower lists to just 5% of the original data, the connectivity rate drops for the star cascade only 2% and for the complex cascade by 20%. The reason for this is that most users actually don't exert much influence, while multiple diffusion paths compensate for the lost social connections in complex structures.

For case (2), we removed randomly chosen messages in order to investigate the impact on the reconstruction rates. According to the lower part of Table 1 connectivity rate drops significantly when removing random retweets: we encounter a decrease of more than 20% for star cascades and 30% for complex cascade when keeping 75% of messages. That means that connections missing due to absent messages cannot be compensated.

Overall, missing messages due to rate limiting results in worse results than missing social graph data. As a result, retrieval of messages is more important in keeping the cascade connected than social links, which also supports our crawling approach focusing on users with high activity or a large amount of followers.



(a) Retweet Count (b) Diameter

(c) Size vs Root Out-degree (d) Size vs non-followers tweets

Figure 3: Cascades' Properties

5.3 Cascade Properties

A preliminary analysis of properties of information cascades is presented in this section.

In a first step, we studied basic properties of cascades. We consider cascades with more than 10 retweets since we can find more complex structures on them while small cascades have been already studied [10, 3]. We can observe a skewed distribution for retweet counts with the biggest cascade containing more than 60K retweets (Figure 3 a). The diameter shows that cascades tend to be deep, with a mean value of diameter 4 (Figure 3 b), contradicting previous studies even for the big cascades [16, 12]. Diameter values up to 18 are observed, indicating that information is being propagated to a large audience much beyond the root's followers. This has an impact on cascades' shapes, that demonstrate complex structures more often than star structures. Another observation is that cascades with diameter 1 are observed with the same frequency as the ones with diameter 5. That confirms again our observations that cascades are more deep than swallow, paving the way for complex analyses.

Since we unravel big and complex cascades with long paths, we studied the role of the root in originating such cascades. Is the root highly influential or cascades tend to be big and deep due to users who forward the tweet of root? Figure 3 c) and d) show that cascade size is correlated both with the direct followers of the root who retweeted (root out-degree) and with the non followers who retweeted. As a result, both the influence of the root and users who forward further the message in combination yield big cascades. On the contrary, there is no correlation (Correlation Coefficient = 0.14) between the size of the cascade and the number of followers of root. Number of followers of a user is not informative of his influence. In this case, results are consistent with previous results in [16, 4]

5.4 Impact of Different Influence Models

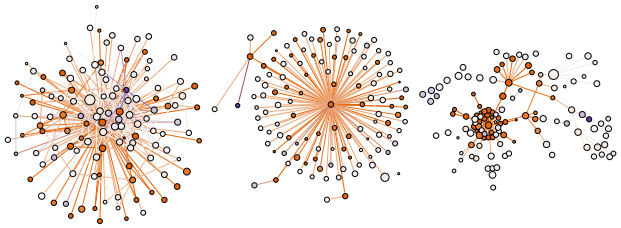
Since we get very good reconstruction results in cases of gradually removing influence edges, we concluded and observed that there exist multiple influence paths, hence a large number of possible influences. We study (1) how many of the cascades are actually affected by different influence models, and (2) how cascades metrics and properties are affected, based on a concrete example.

For 10% of the nodes we can observe on average more than 3 influencers, while 20% of the cascades maintain an average number of influencers greater than one. As a result, different influence

	100%		75%		50%		25%		15%		10%		5%	
	S	C	S	C	S	C	S	C	S	C	S	C	S	C
Connectivity Rate	90,9	87,74	90,77	87,26	90,54	85,59	89,82	81,71	89,33	79,02	88,87	76,35	88,35	71,03

	100%		96,8%		93,7%		87,5%		75%	
	S	C	S	C	S	C	S	C	S	C
Connectivity Rate	90.90	87.74	88.02	80.38	85.18	77.05	79.04	70.75	67.76	58.33

Table 1: Impact of missing followers (upper table) and missing messages (lower table) - Star shape (S), Complex structure (C) - 1000 retweets



(a) Complete (b) Earliest (c) Latest

Figure 4: Impact of different influence models

models described in Section 2 do matter for 20% of the cascades. This may underestimate the real results, since the datasets are incomplete, in particular possible influence edges may be missing.

Different influence models have a big impact on cascade metrics and properties. First, paths lengths are affected. Since we simplify multiple influence paths to one according to a specific model, the number of edges is decreasing while path lengths are increasing. Also, the temporal distribution of edges changes according to the model: the earlier influencer model produces edges closer to the root's timestamp, while latest influencer favors a more stretched distribution of late retweets. In addition, the out-degree of users in the cascade changes according to different model.

In Figure 4 three models were used to reconstruct the same cascade with size 124 nodes. In the complete case, all influence paths were considered, while in the earlier and latest influencer models only one edge was selected according to models in Section 3. Node colour signifies temporal behaviour: the more red, the smaller the temporal distance from root, the more grey-blue the greater the temporal distance from root. Node size varies in log scale according to the number of followers a user has, showing how big is the audience that this user can potentially influence. We can observe that the structure and paths change dramatically for different models. For the complete reconstruction and the earliest-retweet model, the diameter is 3 while for the latest-retweet model, the diameter becomes 11. This can be explained by the fact that since we keep the latest influencer on time, we choose the longest path. Moreover, that explains the greater number of grey-blue nodes in the last model (greater temporal distance from root).

6. CONCLUSION AND FUTURE WORK

In this paper, we present the first steps towards the real-time analysis of information diffusion and user interaction in social media. We introduce models and methods to reconstruct information cascades over real-life Twitter data. We showed that such a reconstruction is feasible and social links play predominant role in information diffusion. Noisy data does have an impact, but we understand which aspects are most critical and we work on ways to overcome these limitations. A preliminary analysis of these cascades shows that they exhibit significantly more complexity as previous studies indicated, paving the way for richer studies. In the future, we plan to cover a much broader range of datasets and extend our evaluations to other kinds of information propagation other than retweets. Also, we plan to build lightweight models in order to infer missing messages and social links. Lastly, we will target many engineering parts of the system, since we cannot yet sustain the scale and performance requirements needed for a full social network.

7. REFERENCES

[1] Storm. http://storm-project.net/.

[2] F. Alvanaki et al. See what's enblogue: Real-time emergent topic identification in social media. In *EDBT*, pages 336–347, 2012.

[3] E. Bakshy et al. Everyone's an influencer: Quantifying influence on Twitter. In *WSDM*, pages 65–74, 2011.

[4] M. Cha et al. Measuring user influence in Twitter: The million follower fallacy. In *ICWSM*, 2010.

[5] P. Cogan et al. Reconstruction and analysis of Twitter conversation graphs. In *HotSocial '12*, pages 25–31, New York, NY, USA, 2012. ACM.

[6] A. Dotey et al. Information diffussion in social media, 2011. http://snap.stanford.edu/class/cs224w-2011/proj/mrom_Finalwriteup_v1.pdf.

[7] J. Goldenberg et al. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223, 2001.

[8] M. Granovetter. Threshold models of collective behavior. *American journal of sociology*, pages 1420–1443, 1978.

[9] A. Guille et al. Information diffusion in online social networks: A survey. *SIGMOD Record*, 42(2):17, 2013.

[10] C. Hui et al. Information cascades in social media in response to a crisis: A preliminary model and a case study. In *WWW (Companion Volume)*, pages 653–656, 2012.

[11] G. Kossinets. Effects of missing data in social networks. *Social Networks*, 28(3):247 – 268, 2006.

[12] H. Kwak et al. What is Twitter, a social network or a news media? In *WWW*, pages 591–600, 2010.

[13] J. Leskovec et al. Patterns of cascading behavior in large blog graphs. In *SDM*, 2007.

[14] M. Mathioudakis and N. Koudas. Twittermonitor: Trend detection over the Twitter stream. In *SIGMOD Conference*, pages 1155–1158, 2010.

[15] A. Tumasjan et al. Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Association for the Advancement of Artificial Intelligence*, 2010.

[16] Z. Zhou et al. Information resonance on Twitter: Watching Iran. In *Social Media Analytics*, SOMA '10, pages 123–131, 2010.

[17] B. Zong et al. Inferring the Underlying Structure of Information Cascades. In *ICDM*, 2012.