Article

Interaction strength analysis to model retweet cascade graphs

Paola Zola [1,*], **Guglielmo Cola** [1], **Michele Mazza** [1,2] **and Maurizio Tesconi** [1]

1 Institute for Informatics and Telematics (IIT) of the National Research Council of Italy (CNR), Pisa, Italy
2 Dept. of Information Engineering, University of Pisa, Italy
* Correspondence: paola.zola@iit.cnr.it

Abstract: Tracking information diffusion is a non-trivial task and it has been widely studied across different domains and platforms. The advent of social media has led to even more challenges, given the higher speed of information propagation and the growing impact of social bots and anomalous accounts. Nevertheless, it is crucial to derive a trustworthy information diffusion graph, which is capable of highlighting the importance of specific nodes in spreading the original message. The paper introduces the interaction strength, a novel metric to model retweet cascade graphs by exploring users' interactions. Initial findings show the soundness of the approaches based on this new metric with respect to the state-of-the-art model, and its ability to generate a denser graph, revealing crucial nodes that participated in the retweet propagation. Reliable retweet graph generation will enable a better understanding of the diffusion path of a specific tweet.

Keywords: social media; network analysis; interaction strength; retweet graph; retweet cascade

---

1. Introduction

In recent years, the explosion of Web 2.0, blogs, microblogs and online social media dramatically changed information consumption and spreading. A recent survey reveals that, for the first time in the history, United States people consume more news from social media than newspaper[1]. Thus, tracking the information diffusion, especially on online communities, is a very important step which is useful for many application such as early warning systems, social bot and communities detection, user location prediction, financial recommendations, marketing campaign effectiveness, political mobilization and protests, etc [1–4].

Among online communities, social media represent a preferable channel for information diffusion and with more than 330 million of monthly active users, Twitter is one of the most used social media which is often considered as an information network [5,6]. Twitter offers four possible actions to express interest in a specific content: favorite, reply, quote and retweet. Replying or liking a tweet does not involve the spread of the content, whereas quotes and retweets are actions used to share information with a wider audience. However, quoting or retweeting a message may indicate a different user behavior. A retweet is often considered an endorsement, i.e. the user supports the original tweet content, whereas quoting may be done in order to express a different idea [7].

In order to understand the connections among users it is important to consider not only their social network but also the way they interact with information, especially through retweets [5,6]. Thanks to the Twitter API service, it is possible to collect a huge amount of information regarding tweets, accounts, users timelines and social networks (i.e., following and followers). However, the Twitter API does not provide complete information about retweets and their propagation path. More precisely, the only information carried by a retweet is the original author of the tweet, whereas possible intermediate steps (i.e., retweeting from a retweeter) are lost.

---

1 https://www.cnbc.com/2018/12/10/social-media-more-popular-than-newspapers-for-news-pew.html

To estimate retweet cascade graphs, previous studies typically adopted strategies based on social network information (i.e., friends and followers) in conjunction with temporal information. These studies exploited the fact that users tend to interact more often with newer tweets [7], and thus a user is more likely to retweet the last friend who retweeted a content. However, this approach is no longer a reliable way of estimating retweet graphs, as Twitter does not show contents based on a simple reverse chronological order, but according to user interests, trending topics, and interactions[2].

Another factor that needs to be considered is the time required to fetch all the required social network information. Due to the Twitter API rate limits, the time required to collect the list of friends and followers is six times greater with respect to downloading the user timeline[3], on average.

1.1. Research Objective and contributions

Following the limitations of the existing approach to generate retweet cascade graphs, in this paper we introduce the concept of *interaction strength* (IS), a novel metric that describes the strength of the link between two accounts in terms of reciprocal interactions, including quotes, replies, and retweets. Based on this new metric, we propose two novel approaches to estimate retweet cascade graphs. The first approach is the *interaction strength-based network* (ISN), where the graph is constructed by maximizing the overall IS value computed for each pair of accounts. The friend list is used only for accounts that did not interact with other users, and for which it is not possible to find the IS value. The second approach is called *interaction strength-based network with author's followers evaluation* (ISN-AF). This approach is similar to ISN, but the author's follower list is exploited to generate the first level of the retweet cascade graph. Indeed, on Twitter a follower of the author of the original tweet (i.e., the root of the graph) can only retweet from the author himself/herself. It should be highlighted that both ISN and ISN-AF require to fetch the friend list for a limited number of accounts, whereas in the traditional approach it is necessary to fetch the friend list for each retweeter in the graph.
Models aimed at constructing the retweet cascade graph suffer from the absence of a ground truth to check the correctness of the method. Nevertheless, following [8], we evaluated the proposed approaches in terms of network graph strength, considering several metrics.

The proposed work novelty and contribution can be summarized as follows:

- we introduce the concept of *interaction strength* (IS), a metric that indicates the strength of the link between two users
- we propose two novel approaches based on IS to generate retweet cascade graphs – ISN and ISN-AF;
- ISN aims to maximize IS values for each pair of nodes in the graph;
- ISN-AF is similar to ISN, but the first level in the retweet cascade graph is based on the list of followers of the root user;
- both of the proposed approaches are mainly based on information contained in the users' timelines, which can be conveniently retrieved through the free Twitter API service (compared to fetching the list of friends for each node as in the traditional approach, which is a substantially more time-consuming task);
- the source code is freely available on GitHub [4].

In Section 2 we provide a detailed overview on previous work about information propagation and retweet cascade prediction. Then, in Section 3 the proposed IS concept is described jointly with the ISN model, while, in Section 4 we report the description of the ISN-AF approach. Section 6 details the dataset and Section 7 reports the analysis of the results related to the IS metrics while the comparison

---

2 https://blog.hootsuite.com/twitter-algorithm/
3 https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/api-reference/get-followers-ids
4 https://github.com/paolazola/Interaction-strength-analysis-to-model-retweet-cascade-graphs

among the proposed ISN and ISN-AF with respect to the baseline are in Section 8. Finally, Section 9 concludes the paper.

2. Related Work

In Table 1 we report a summary, in chronological order, of previous work about information propagation on Twitter. As Table 1 shows, the majority of works has focused on predicting retweet engagement (REP) in terms of total number of retweets. Other studies conducted a joint analysis between tweet information cascade (TIC) and REP, assuming that the retweet chains can be deducted from tweet content.

Table 1. Summary of related works

Study	Target[a]	Dataset[b]	Dataset[c] Size	Dataset Collected in	Topic Features	Text Features	Time Variable	Users Features	Users Interactions	Social Network	Location Features	Users Behaviour	Model[d]
Szabo et al. [9]	REP	YT, D	YT: 7K, D: 850K U	2007-08	-	-	X	-	-				LR
Yang et al. [10]	TIC-REP	TW	-	-	X	X	X	-	-	-	-	X	F+EM
Cogan [11]	TIC	TW	33K T	2012	-	-	-	-	-	X	-	-	RCM
Comarela et al. [7]	RB	TW	54M U	2006-09	-	X	X	X	-	-	-	X	SVM, NB
Yang et al. [12]	RB	TW	22M T	2009	X	-	X	-	-	-	-	-	CHR
Remy et al. [13]	TIC	TW	362M T	2011	-	-	-	-	-	X	-	-	PL
Zaman et al. [14]	TIC-REP	TW	52	-	-	-	X	X	-	-	-	-	HB
Taxidou et al. [8]	TIC	TW	11M T	2012	-	-	X	-	-	X	-	-	
Pramanik et al. [15]	TIC	TW	55K	-	X	-	X	-	-	X	-	-	H
Yu et al. [16]	TIC-REP	TWB	320M U	2011	-	-	X	X	-	-	-	X	NEWER
Zhao et al. [17]	REP	TW	3.2B	2011	-	-	X	X	-	-	-	-	SEISMIC
Gao et al. [18]	TIC-REP	SW	164	-	-	-	X	-	-	-	-	-	RPP
Kobashy et al. [19]	TIC-REP	TW	166K	2011	-	-	X	X	-	X	-	-	TiDeH
Rodrigues et al. [20]	TIC	TW	17K	2013	-	X	X	-	-	X	X	-	GetMove
Cao et al. [21]	REP	SW, PC	50K T, 35K P	2016	-	-	X	-	X	-	-	-	DH
Zhou et al. [22]	TIC-RB	SW	69.4M	2013-14	X	-	X	X	-	X	-	X	BN
Stai et al.[23]		TW	35K	2014-16	X	-	X	-					EpiM
Bhowmick et al. [24]	TIC-KR	TW	8M T	2015-18	-	-	X	-	-	X	-	-	SmartInf
Chen et al. [25]	REP	TW	20K	2016	-	X	X	-	-	-	-	-	NPP
Wu et al. [26]	KR-TIC	SW	50K M	-	-	-	X	X	X		X		RL2R
This work	TIC	TW	16K T	2020	-	-	X	X	X	X	-	-	W-RCM

[a] **Target** – Key Retwetters (KR), Tweet Information Cascade (TIC), Retweeting Behaviour (RB), Retweet Engagement Prediction (REP)
[b] **Dataset** – Digg (D), Papers' Citations (PC), Sina-Weibo (SW), Twitter (TW), Tencent-Weibo (TWB), YouTube (YT)
[c] **Dataset Size** – Thousand (K), Messages (M), Papers (P), Tweets (T), Users (U), YouTube Videos (YT)
[d] **Model** – Linear Regression (LR), Features (F), Expectation Maximization (EM), Support Vector Machines (SVM), Naive Bayes (NB), Relation base Learning to Rank (RL2R), Retweet Cascade Modelling (RCM), Cox proportional Hazard Regression (CHR), Power Law (PL), Neural Popularity Prediction (NPP), DeepHawakes (DH), SmartInfluencer (SmartInf), Hawkes process (H), Epidemic Model (EpiM), Reinforcement Poisson Process (RPP), Networked Weibull Regression (NEWER), Hierarchical Bayesian approach (HB), 7 Metrics (7M), Self Exciting Point Process (SEISMIC), Bayesian Networks (BN)

For instance, [10,12,14,21] derived the retweet propagation paths assuming that the user's "ScreenName" reported in the text, such as "*RT@ user ScreenName*" indicates the user from whom the current user has read the message. However, as argued by [13] this assumption is mostly inaccurate. Moreover, the user information held in the current Twitter API is only related to the actual user retweeting the message and the original tweet author, thus ignoring intermediate accounts [8]. More reliable retweet information cascades are the ones that merge temporal and social network information [15]. Scholars widely investigated about factors affecting user retweet behaviour (RB), finding that accounts with higher number of followers tend to be retweeted more often. However, there is not agreement on the minimum number of followers needed to be regarded as an "influencer" [6,13]. A relevant study in this field is [8], which analyzed four different options based on followed accounts to derive the possible retweet graph. Since there is no ground truth to compare the possible cascade options, the authors evaluated the options computing two metrics: the Connectivity-Rate and the Root-Fragmented-Rate. Other works proposed the use of additional features such as users' metadata (e.g., number of followers, number of friends, statuses count etc) [16,17], text and topic similarities features [10], location information [20,26] and users tweeting behaviour (e.g., incidence of tweet or retweet in the user activity) [7]. Only two studies [21,26], as far as we know, integrated in the retweeting dynamic analysis the impact of social relationships measured as the intensity of the interactions between two users. However, in both works [21,26], the authors measured the interactions in terms of retweets which are, accordingly to the actual Twitter API information, not completed and miss intermediate steps introducing a bias. Thus, in this paper, we further refer to this bias as *retweet bias*. Nonetheless, as argued by [27], trust between users is an important factor for information dissemination on distributed online social networks.

Therefore, starting from the findings in [27,28] and considering the limitations of the Twitter API service, in this paper we propose a novel approach to generated retweet cascade graphs based on users interaction strength (IS), which is measured taking into account not only retweets, but also quotes and replies.

3. Interaction strength-based network (ISN) approach to generate retweet cascade graphs

An information cascade C is defined as a directed graph $C = (\mathbb{V}, \mathbb{E})$ in which each node $u \in \mathbb{V}$ represents a user u and each edge $(u, y) \in \mathbb{E}$ represents the link from user u to user y. A retweet cascade graph is a class of information cascade characterized by a tree structure where the root node is the author of the original tweet ($root_{author}$), which was posted at time $root_{time}$.

Our purpose is to estimate the retweet information cascade graph using the following information:

- the retweets creation time t_r;
- the interaction strength between each couple $(u, w) \in \mathbb{V}$, which reflects the trust between users;
- the friend lists $L_{\mathbb{F}}$ for the remaining nodes $\mathbb{F} \subset \mathbb{V}$, for which no interactions were found (e.g., $IS_{\mathbb{F}} = Null$).

A link in the cascade between any two nodes in $\mathbb{V}$ has to meet the following condition:

$$\begin{aligned} \exists \quad E(u,w) \in \mathbb{E} \quad \forall \quad u,w \in \mathbb{V} \rightarrow t_u > t_w \quad \wedge \quad \exists \quad IS_{u,w} \\ s.t. \quad IS_{u,w} = \max(IS_{\mathbb{V} \setminus \{u\}}) \end{aligned} \tag{1}$$

In other words, the user u is connected to the user w if w retweeted the message before u and if the IS of u with respect to w is the maximum of all the IS among user u and the other accounts that retweeted before u. The procedure to find the IS value is described in the next subsection. Whenever the user $u \in \mathbb{V}$ has no interactions with any other accounts, the proposed method adopts the approach based on social networks [8] collecting user u's friends list L_u such that:

$$\begin{aligned} \exists \quad E(u,w) \in \mathbb{E} \quad \forall \quad u,w \in \mathbb{V} \rightarrow w \in L_u \quad \wedge \\ t_w = min(t_i - t_u) \quad \forall \quad i \in L_u \cap \mathbb{V} \setminus \{u\} \quad \wedge \quad t_i > t_u \end{aligned} \tag{2}$$

If there is no information about interactions, the node u is connected to the user in his/her friend list L_u that retweeted the tweet $root_{author}$ before u, using an inverse chronological order, i.e., minimizing the difference between the user u posting time t_u and t_i, where i belongs to the set of users in L_u. In the following, Section 3.1 and Section 3.2 describe the steps performed to find the interaction strength (IS) and to evaluate different sets of possible weights. Then, Section 3.3 explains how to connect remaining users according to social network information.

3.1. Twitter Users Interaction strength

We here describe the procedure that exploits the interactions to generate the cascade graph. For a user u, it is first downloaded its timeline by means of the Twitter API. Then, it is found its absolute interaction strength $AbsoluteIS_{u,w}$ with respect to any node w that retweeted before u. More precisely, to derive the $AbsoluteIS_{u,w}$ the following information is required:

- number of quotes $Q_{u,w}$ that user u expressed to the node w;
- number of replies $RP_{u,w}$ that user u did to w;
- number of retweets $RT_{u,w}$ that user u did to w.

The absolute interaction strength between u and w is defined as:

$$AbsoluteIS_{u,w} = Q_{u,w} + RP_{u,w} + RT_{u,w}.$$

The $AbsoluteIS$ is found only with respect to nodes that retweeted before u, and we also discarded interactions that happened after the root tweet creation ($root_{time}$) as well as those that are too old to represent an ongoing trust relationship (trust relations tend to change over the years). To be more precise, we discarded interactions that occurred more than two years before the $root_{time}$. We relied on $root_{time}$ as threshold date for all the users relations involved in the cascade C since, in general, the tweet information cascade lifetime is short and concentrated in proximity of the tweet creation date. Figure 1 depicts the temporal dynamics of the retweets after the respective $root_{time}$, showing a decreasing trend, as the highest number of interactions occur soon after $root_{time}$ (the original tweet creation date).

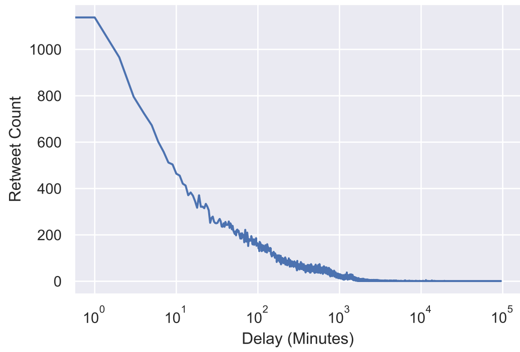


Figure 1. Number of retweets received after the original tweet creation date.

After, a weighted IS value $WeightedIS_{u,w}$ is found, by assigning different weights to quotes (w_q), replies ($w_r p$), and retweets ($w_r t$):

$$WeightedIS_{u,w} = w_q * Q_{u,w} + w_{rp} * RP_{u,w} + w_{rt} * RT_{u,w}.$$

Finally, the $IS_{u,w}$ value is found dividing $WeightedIS_{u,w}$ by the sum of all the weighted IS values for the user u. In this way, we obtained IS values for u that range within $[0,1]$, and are thus easier to evaluate and compare among retweets cascades. For instance, if $IS_{u,w}$ is 1, this means that w is the only user that retweeted before u and with whom u has interacted. The next Section describes the grid-search procedure adopted to derive the optimal weights (w_q, w_{rp}, w_{rt}).

3.2. Interaction weights for retweet cascade graph

Table 2. Weight sets – Guideline setups

Set nr	Retweet weight	Quote weight	Reply weight
0.0	1	1	1
0.1	1	0	0
0.2	0	1	0
0.3	0	0	1

To evaluate different weight values for the three interactions we conducted experiments based on a grid search approach with 15 different sets of weights. Four sets are named *Guideline setups* and are shown in Table 2: in the first guideline setup (weight Set 0.0) all the weights are set to 1 and thus each interaction has the same impact on the approach; the remaining three guideline setups (Sets 0.1, 0.2, 0.3) have only one weight set to 1, whereas the others are set to 0 to evaluate the impact of each interaction separately. Finally, 11 *Experimental setups* are shown in Table 3, which were used to identify the best weight sets for the proposed approaches.

After defining the set of weights, it is possible to derive the $WeightedIS_{u,w}$ for each pair $(u,w) \in \mathbb{V}$. Thus, for each user u we can find a collection of accounts $\mathbb{O} = \mathbb{V} \setminus \{u\}$ that retweeted before u and that interacted with u, such that $\exists k \in \mathbb{O} \mid WeightedIS_{u,k} \neq 0$. As mentioned before, the $IS_{u,w}$ value is found dividing by the highest weighted IS value for u.

In Table 4 we report a toy example of $AbsoluteIS$ and $WeightedIS$ for a user u with respect to some other users (w,h,j,y), adopting the experiment Set number #1 from Table 3. The $AbsoluteIS$, simply

Table 3. Weight sets – Experimental setups

Set nr	Retweet weight	Quote weight	Reply weight
1	0.05	0.50	1.00
2	0.05	1.00	0.50
3	0.50	0.05	1.00
4	0.50	1.00	0.05
5	1.00	0.05	0.50
6	1.00	0.50	0.05
7	0.25	0.75	1.00
8	0.25	1.00	0.75
9	0.30	1.00	0.80
10	0.15	0.95	0.65
11	0.35	1.00	0.70

computed as sum of the raw users interactions, is then translated to the $WeightedIS$ by multiplying each interaction according to the adopted weight set. Then, in Table 5, IS values are found dividing by the sum of the $weightedIS$ values for u, that in this example is $17,1 + 7,85 + 3,05 + 2,50 = 30,5$. The highest IS value is obtained by w, and thus in the cascade u can be linked to w. The toy example gives an idea of how the weights can change the initial interaction strength; in fact considering only the $AbsoluteIS$, user y has an overall absolute interaction with u of 50, whereas for user w the absolute interaction with u is only 24. However, in this example, the relatively low weight for retweets leads to a higher IS value for w.

Table 4. Toy example of user u Absolute and Weighted IS

Interaction With	Interaction Type	Interaction Strength	
		Asbsolute	Weighted
User W	$RT_{u,w}$	2	0.10
	$Q_{u,w}$	10	5.00
	$RP_{u,w}$	12	12.00
	Sum	*24*	*17.10*
User H	$RT_{u,h}$	7	0.35
	$Q_{u,h}$	15	7.50
	$RP_{u,h}$	0	0.00
	Sum	*22*	*7.85*
User J	$RT_{u,j}$	1	0.05
	$Q_{u,j}$	2	1.00
	$RP_{u,j}$	2	2.00
	Sum	*5*	*3.05*
User Y	$RT_{u,y}$	50	2.50
	$Q_{u,y}$	0	0.00
	$RP_{u,y}$	0	0.00
	Sum	*50*	*2.50*

Table 5. Toy example of user u IS

	AbsoluteIS	WeightedIS	Interaction Strength
User W	24	17.10	0.56
User H	22	7.85	0.25
User J	5	3.05	0.10
User Y	50	2.5	0.08

3.3. Users without interactions and sparse nodes

As a complementary step, when there are no available interactions by a user u, and thus no IS values between u and any other user, we attempt to find a link from the u to another user in the cascade according to the state-of-the-art method based on social networks. More precisely, we collect the user friend list by using the Twitter API and every user's friend that has retweeted at an earlier point in time is considered as a potential influencer [6,13,29]. To identify the influencer that more likely spread the tweet to the user u we consider the most recent influencer, i.e., u is linked to the last friend that retweeted the message. Users that still remain without an edge after this second step are denoted as *Sparse Nodes* (SN).

4. Alternative approach: information strength-based network with author's followers evaluation (ISN-AF)

As a further version of the model based on the IS concept (ISN approach) we also propose a modified algorithm that first exploits the tweet's author followers network. In fact, Twitter preferably shows original contents, and thus if a user directly follows the tweet's author, he/she will retweet the original tweet without seeing the retweets from intermediate nodes. To exploit this Twitter feature, the ISN-AF approach first explores if a retweeter belongs to the tweet's author followers network: followers are linked directly to the root in the cascade graph. For remaining nodes, ISN-AF uses the same approach as in the ISN model, namely IS analysis and friends network for sparse nodes. This algorithm version allows to reduce the computation costs related to the IS step but, at the same time, requires additional information which is the tweet's author followers list, which can be time consuming if the account has a lot of followers. Moreover, this ISN-AF version suffers from the limitation reported in Section 1 about the possible divergence in terms of followers relationship between the date in which the retweet has been done and the time when the followers list is fetched.

5. Evaluation metrics for retweet cascade validation

Evaluating the models ability in deriving the retweet cascade graph is not a trivial task. The absence of ground truth information prevents the use of standard evaluation metrics such as accuracy. Following [8] we evaluate the retweet information graph considering all the cascade forest, i.e. including the unconnected components, and computing the following metrics:

1. Cascade Average Strength (CAS): given the IS assigned to each edge $(u, y) \in \mathbb{E}$ we derive the CAS as the average of the maximum IS between each pair of edges in $\mathbb{E}$ such as:

$$CAS = \frac{\sum(\max(IS_{u,y} \forall u, y \in \mathbb{E}))}{|\mathbb{E}|} \tag{3}$$

2. Connected Components Count (CCC): it return the number of the connected components in the network. A connected component is a subgraph in which any two vertices $v \in V$ are connected to each other by paths.
3. Root Fan Ratio (RFR): it assesses whether there is a path to the $root_{author}$ from every other user. In other words, it measures the percentage of nodes directly connected to the root. In the ISN-AF model, the RFR asses the percentage of $root_{author}$ followers.
4. Giant Component Size (GCS): the size, expressed in percentage of the cascade nodes, of the nodes present in the giant component (GC) which is the connected component with biggest size. The GCS is computed as follow:
$$GCS = \frac{|u \in \mathbb{V} \longrightarrow u \in GC|}{|\mathbb{V}|} \tag{4}$$
5. Global Reaching Centrality (GRC): it is the average over all nodes of the difference between the node local reaching centrality and the greatest local reaching centrality of any node in the graph.

The local reaching centrality, $C_R(i)$, of node i is the proportion of all nodes in the graph that can be reached from node i via outgoing edges [30].

$$GRC = \frac{\sum_{i \in \mathbb{V}} (C_R^{max} - C_R(i))}{|\mathbb{V}| - 1} \tag{5}$$

6. Sparse Node Incidence (SNI): it measures the incidence (in percentage) of sparse nodes (i.e., nodes without links) with respect to the total number of nodes in the cascade.

Those metrics are computed in order to identify the best IS weights set among the proposed ones in Table 3. Moreover, the metrics are evaluated in order to compare the proposed ISN and ISN-AF methods with respect to the baseline approach proposed by [8].

6. Dataset

For the analysis, we collected a dataset of Tweets from the 1^{st} January 2020 to the 31^{st} March 2020 written in Italian and related to politic topic. The dataset was collected using the freely available Twitter streaming API service to catch all the retweets related to the original contents. Table 6 reports a summary of dataset composition and, in order to reduce the computational costs, we conduct the experiments on a smaller sample of the data randomly chosen. The sample of tweets used to test the 15 weights set is composed by 16,304 tweets resulting in 244,560 cascades with more than 1.6 million nodes.

Table 6. Dataset description

	Full Dataset	**Sampled Dataset**
Tweets Count	506,147	16,304
Unique Users Count	102,468	41,592
Retweets Count	683,189	112,188

7. IS weights evaluation

We here report the results considering the IS weight sets investigated during the grid search procedure. The aim of this Section is to evaluate the best setup in order to maximize the IS value and generate retweets cascades with high CAS. This step of the analysis is common for both ISN and ISN-AF models which will be evaluated in Section 8. Hence, this preliminary analysis is conducted without considering friend networks, in order to remove external factors in the weights evaluation.

Before investigating the properties of the proposed approach, we aim to analyze the relations, in terms of correlation, among the considered metrics. Figure 2 shows that the CAS values, which represent the average cascade strength, are positively correlated to the RFR. In fact, the higher is the concentration of nodes around the $root_{author}$, the higher is the IS between each node and the *root*. Interestingly, there is a high negative correlation with the depth variable, which corresponds to the number of levels in the cascade tree. In fact, Figure 2 shows that the CAS is strongly negatively influenced by the depth, whereas the node count variable has a smaller incidence. It follows that even big cascades with few levels might achieve a high CAS score. At the same time, the higher is the GRC, the smaller is the CAS. Recalling that the GRC provides an average measure of all cascade's nodes centrality, when the nodes in the cascade are centrals and might be reached by several path, the strength of the cascade decreases since multiple edges exist among users and thus, the selected relations (maximum IS criterion) are weaker.

7.1. Experiments results on the entire cascade sample

Starting from the guideline sets results in Table 7, we got that the set 0.0, which describes the proposed model when all the interactions have the same weight, reaches an average CAS of 87% with

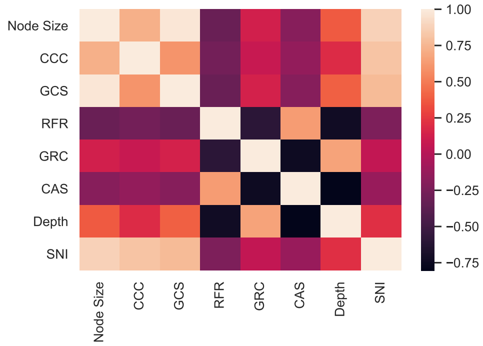


Figure 2. Correlation Matrix among the cascades' evaluation metrics

the 91% of edges passing from the $root_{author}$ (RFR) and less than the 2% of SNI. A growing SNI is visible for the guideline sets 0.1, 0.3 and 0.2, especially for the last two which are characterized by and average CAS of 84%. Guideline set 0.1 is, in general, preferable than sets 0.2 and 0.3 in terms of CAS.

Table 7. Average metrics for guidelines setup

Class	Set	Edge Count	CCC	GCS (%)	RFR (%)	GRC (%)	CAS (%)	Depth	SNI (%)
Guideline	0.0	5.88	1.08	6.25	93.81	6.19	$87.06^{\dagger,\diamond}$	0.53	1.98
Guideline	0.1	5.88	1.08	6.29	93.75	6.18	$85.39^{\star,\diamond,\bullet}$	0.53	2.05
Guideline	0.2	5.89	1.08	6.27	93.68	6.15	$84.33^{\dagger}$	0.53	2.13
Guideline	0.3	5.89	1.08	6.29	93.68	6.15	$84.32^{\star,\dagger}$	0.53	2.11

$^{\star,\dagger,\diamond,\bullet}$ – Statistically significant (p-value < 0.05) under a pairwise comparison when compared with the Guideline Sets: 0.0 (*), 0.1 (†), 0.2 ($^\diamond$), 0.3 ($^\bullet$)

Table 8. Average metrics for each weight set in the experimental setups, computed on all the dataset

Set	Edge Count	CCC	GCS (%)	RFR (%)	GRC (%)	CAS (%)	Depth	SNI (%)
1	5.88	1.08	6.27	93.61	6.12	$83.53^{\star,\diamond,\bullet}$	0.53	2.04
2	5.89	1.08	6.25	93.59	6.12	$82.93^{\star,\diamond,\bullet}$	0.53	2.01
3	5.89	1.08	6.29	93.77	6.17	$86.32^{\dagger,\bullet}$	0.53	2.08
4	5.89	1.08	6.25	93.81	6.19	$87.23^{\diamond,\bullet}$	0.52	1.98
5	5.80	1.08	6.20	93.77	6.06	$84.63^{\star,\diamond}$	0.52	2.03
6	5.80	1.08	6.20	93.74	6.06	$84.21^{\dagger,\diamond}$	0.52	2.04
7	5.80	1.08	6.22	93.78	6.07	$84.80^{\star,\dagger,\diamond,\bullet}$	0.52	2.11
8	5.88	1.08	6.28	93.62	6.15	$83.45^{\star,\dagger,\diamond,\bullet}$	0.53	1.98
9	5.88	1.08	6.27	93.74	6.17	$85.57^{\dagger,\diamond,\bullet}$	0.53	2.06
10	5.88	1.08	6.25	93.77	6.19	$85.90^{\dagger,\diamond,\bullet}$	0.53	2.02
11	5.88	1.08	6.31	93.79	6.18	$86.63^{\dagger,\diamond,\bullet}$	0.53	2.00

$^{\star,\dagger,\diamond,\bullet}$ – Statistically significant (p-value < 0.05) under a pairwise comparison when compared with the Guideline Sets: 0.0 (*), 0.1 (†), 0.2 ($^\diamond$), 0.3 ($^\bullet$) reported in Table 7

Concerning the analysis of the full sample of cascades computed, Table 8 reports the average metrics for each set of weights in the experimental setups. In general, all the proposed experiments reach a valuable CAS which is comprised between the 83% and 87%. Lowest CAS values (around 83%) are achieved by sets 1, 2 and 8 which are also characterized by a smaller retweet weight (0.05, 0.05 and 0.25 respectively). In general, the proposed approach limits the incidence of sparse node (SNI), which is around 2%, with an average cascade depth below 1. However, these results are computed on the entire sample of 16,304 tweets, thus including cascades with few retweets that are characterized by a CAS value of one.

Figure 3 plots on the x-axis the cascade depth while on y-axis the CAS scores achieved by each cascade. This shows the strong negative correlation between CAS and Depth. The negative correlation is common among all the weights sets, but it shows different behaviour according to the incidence of retweets, quotes and replies. For instance, cascades with a higher retweet impact show a more linear decay without drops (sets $\{0.0, 5\}$), but a more interesting evidence is about quotes and replies effects. In fact, sets with identical retweet weight but with quote importance smaller than reply exhibit more variable dynamic with drops for deeper cascades (e.g., set 0.2 versus set 0.3 and set 3 versus set 4).

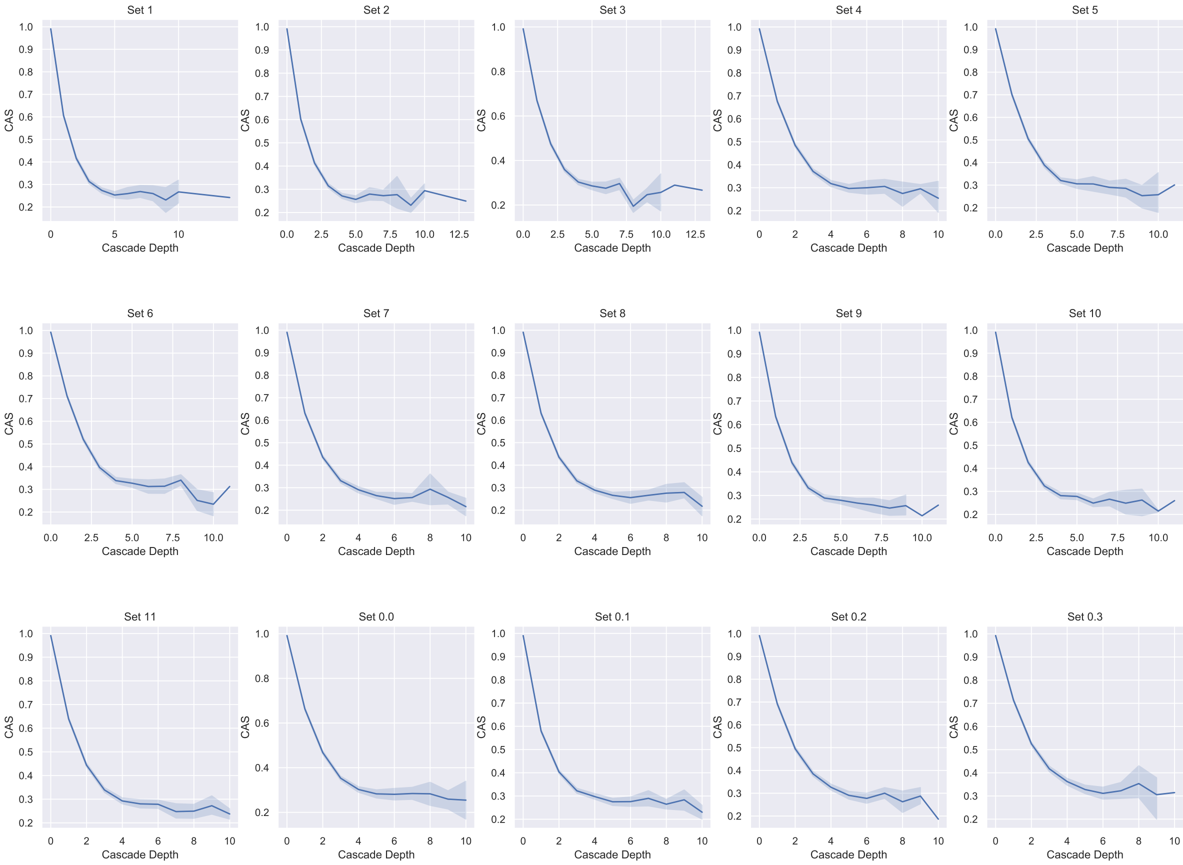


Figure 3. CAS and cascade Depth

Table 9 reports the results of the analysis focused on deep cascades (cascades with more than 6 levels) in order to compare the incidence of quotes and replies in maximising the CAS. It emerges that weight sets with higher quotes incidence are able to achieve higher CAS values (e.g., set 0.3 for group a or set 4 in group b), but, at the same time, when the relative difference between quote and reply weights is high, it also leads to a higher standard deviation. For example, for groups a and b, the difference between quotes and replies weights is high (respectively equal to 1 and 0.95), thus generating a higher average CAS for sets 0.3 and 4, but involving a major variability in the results. It follows that, from Table 9, to limit the CAS decrease and ensure the absence of drops when the cascades are characterized by a high number of levels, a solution is to overweight quote interaction limiting the relative difference with reply interaction. In fact, Table 9 shows that the set able to minimize the standard deviation is the number 11, which also ensures a limited retweet weight (0.35) with respect to the quote and reply weight (1.00 and 0.70 respectively).

7.2. Experiments results on cascades with at least 5 nodes

In terms of absolute CAS results, it is interesting to perform a more detailed analysis removing cascades with few nodes. Focusing on the graphs of the 90^{th} percentile, which are cascades with more than 5 retweets, Table 10 reports the average of the computed metrics. The first four rows report the

Table 9. CAS comparisons for deep cascades (best values in bold)

Groups	Weight Set	Quotes Weight	Replies Weight	CAS	
				Mean	St Dev
a	0.2	0.00	1.00	0.28	0.06
	0.3	1.00	0.00	**0.32***	0.08*
b	3	0.05	1.00	0.28	0.07
	4	1.00	0.05	0.30*	0.08
c	5	0.05	0.50	0.29	0.07
	6	0.50	0.05	0.31	0.08
d	1	0.50	1.00	0.26	0.07
	2	1.00	0.50	0.27	0.07
e	7	0.75	1.00	0.26	0.06
	8	1.00	0.75	0.26	0.06
-	0.0	1.00	1.00	0.28	0.07
	0.1	1.00	1.00	0.28	0.06
	9	1.00	0.80	0.26	0.06
	10	0.95	0.65	0.25	0.06
	11	1.00	0.70	0.26	**0.05**

* – Statistically significant (p-value < 0.05) under a pairwise comparison when compared with the set within the group

results for the weight sets of the guideline setups. Set 0.0 achieves an average CAS of 47%, while cascades where the incidence of retweets is zero (set 0.1) achieve a CAS of 41%, with a slightly increased SNI (16.96% vs 16.33% for set 0.0). This suggests that quotes and replies have a higher impact on CAS with respect to retweets. In fact for the sets 0.2 and 0.3, where the weight of quotes and replies is set to zero, respectively, the CAS value drops to 38%, with an increased incidence of sparse nodes and a reduction in terms of GCS. Observing the first six experimental setups, which use the same three weight values $(0.05, 0.5, 1)$ assigned differently to the three interactions, the best CAS value is achieved by set number 4, where the priority is assigned to quotes, followed by retweets and replies. Observing the CAS distribution reported in the box plots in Figure 5a, it emerges that the highest median is related to set 6 followed by set 5. However, all the weights sets are characterized by a similar distribution which tends to show a positive asymmetry. In contrast, from the heatmap in Figure 5b sets 5 and 6 are characterized by a higher number of cascades, with CAS values above 0.5. Nonetheless, due to the *retweet bias*, our choice is to prefer weight sets in which the impact of retweets is minimal with respect to the other two interactions.

Therefore, the ideal weight set should give less attention to retweets, medium importance to replies, while quotes should represent the interaction with higher priority. Among the remaining sets $2, 8, 9, 10$ and 11, the maximum CAS is achieved by set 11 (44.7%). This set is also characterized by a smaller number of levels (2.38 levels on average), and a high percentage of edges passing thorough the $root_{author}$ (61%), similar to the guideline set 0.0. Indeed, set 11 shows similarities to the guideline set 0.0 in terms of GCS, RFR, number of edges, but with a statistically significant difference in terms of CAS distribution, which is 2 p.p. (percentage points) lower than set 0.0 but respecting the assumption on retweet bias.

8. ISN, ISN-AF and baseline comparisons

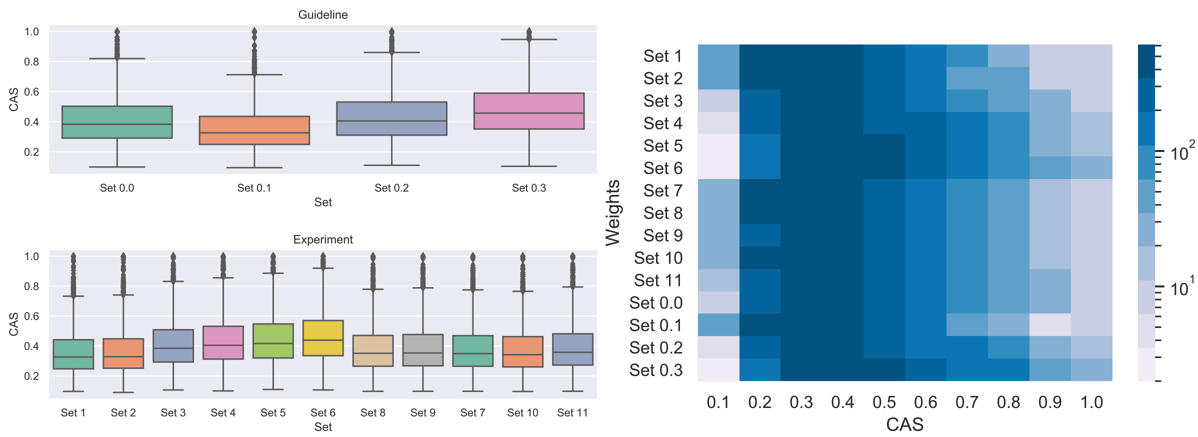
In this Section we compute the retweet cascade graph for all the 16,304 tweets using the proposed ISN approach based on IS and, whenever the IS value is not available, on social network information. Moreover, we also report the results of the ISN-AF approach, where a preliminary check on $root_{author}$ followers is performed to generate the first level of the cascade graph. Results are then compared with a baseline model proposed by Taxidou et al. [8].

Table 10. Average metrics for all the setups (guideline and experimental), computed on the cascades with more than five nodes

Set	Edge Count	CCC	GCS (%)	RFR (%)	GRC (%)	CAS (%)	Depth	SNI (%)
0.0	46.17	1.54	41.96	60.76	32.84	$46.77^{\dagger,\diamond,\bullet}$	2.37	16.33
0.1	46.17	1.56	42.31	60.29	32.71	$41.45^{\star,\diamond,\bullet}$	2.37	16.96
0.2	46.21	1.56	42.09	59.78	32.54	$38.34^{\star,\dagger,\bullet}$	2.40	17.56
0.3	46.21	1.55	42.28	59.80	32.58	$38.46^{\star,\dagger,\diamond}$	2.40	17.43
1	46.17	1.55	42.13	59.37	32.35	$36.09^{\star,\diamond,\bullet}$	2.43	16.81
2	46.21	1.52	41.92	59.43	32.27	$35.90^{\star,\diamond,\bullet}$	2.43	16.54
3	46.21	1.55	42.28	60.51	32.65	$43.65^{\dagger,\diamond,\bullet}$	2.39	17.14
4	46.21	1.52	41.95	60.70	32.82	$48.31^{\star,\dagger,\diamond,\bullet}$	2.36	16.33
5	46.15	1.56	42.16	59.75	32.56	$38.80^{\star,\dagger,\diamond,\bullet}$	2.39	16.96
6	46.15	1.56	42.09	59.58	32.54	$37.71^{\star,\dagger,\diamond,\bullet}$	2.39	17.08
7	46.15	1.56	42.31	59.83	32.58	$39.34^{\star,\dagger,\diamond,\bullet}$	2.39	17.64
8	46.17	1.55	42.20	59.48	32.53	$36.39^{\star,\dagger,\diamond,\bullet}$	2.41	16.35
9	46.17	1.56	42.11	60.28	32.67	$41.55^{\star,\dagger,\diamond,\bullet}$	2.40	17.02
10	46.17	1.54	41.96	60.40	32.78	$43.69^{\star,\dagger,\diamond,\bullet}$	2.37	16.67
11	46.17	1.55	42.46	60.63	32.73	$44.74^{\star,\dagger,\diamond,\bullet}$	2.38	16.47

$^{\star,\dagger,\diamond,\bullet}$ – Statistically significant (p-value < 0.05) under a pairwise comparison when compared with the Guideline sets: 0.0 (*), 0.1 (†), 0.2 ($^\diamond$), 0.3 ($^\bullet$)

Figure 4. CAS on the set of tweets with at least five retweets



(a) CAS Box Plots distribution **(b)** Cascades Heatmap distribution

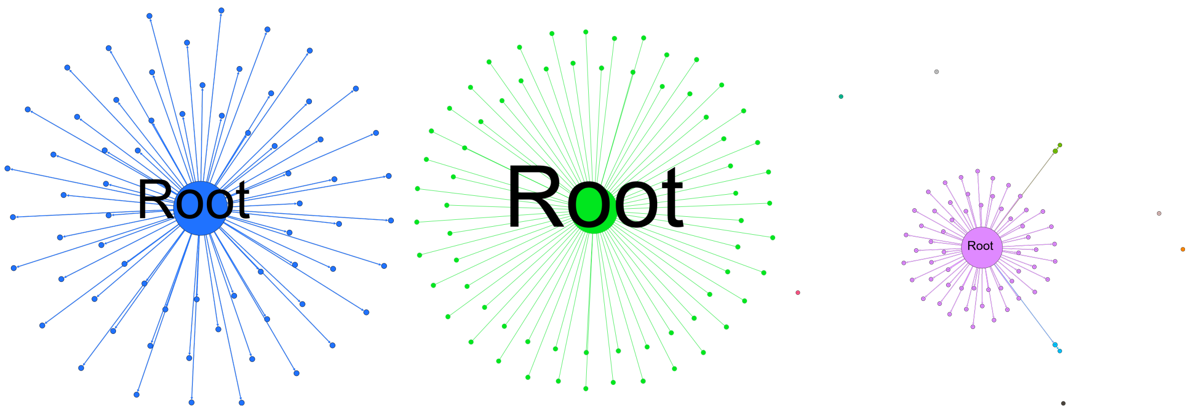
Table 11 reports the metrics for our two approaches and the baseline; the CAS value is only available for the IS and ISN-AF (the baseline model is not based on IS) achieving 85.29% and the 94.7% CAS, respectively. Hence ISN-AF, which exploits the author's followers list, outperforms ISN in terms of CAS. This is due to the probability equal to one assigned to edges connecting nodes to the $root_{author}$, whenever the node is in the author's followers list. Indeed, ISN-AF creates denser cascades around the root node, as indicate by the higher RFR and GRC values. However, as argued in the Section 1, methods based on friends/followers lists have two major limitations: the computational time required to collect the information and, most importantly, the temporal gap between the date in which the retweet is collected and the date when the followers list is fetched. This latter limitation could introduce a form of bias in the cascade.

Comparing the proposed methods with respect to the baseline, a significant difference is apparent regarding the SNI: the two proposed models have an incidence of sparse nodes below 2%, whereas the baseline has an average SNI above 6%. This lower SNI could be linked to the ability of the proposed approach to overcome the limitations of the baseline, as discussed in Section 1 and Section 2. For instance, Twitter accounts might have a private network, which is not accessible, thus the baseline approach based has to classify these node as sparse, whereas the proposed approaches may still be able to find some interactions and find a link for these node in the graph. This ability of the proposed approaches is also confirmed by edge count, which is grater for ISN and ISN-AF with respect to the baseline approach. Moreover, the missing edges for the baseline involve a higher number of isolated connected components, which leads to higher CCC values. At the same time, the major incidence of sparse connected components reduces the giant component size, which is, in general, denoted by the $root_{author}$ component. In fact, the GCS value for the baseline is 4.99%, while it reaches 6.65% for ISN and 6.90% for ISN-AF.

Table 11. Comparison between the two proposed approaches (with the best weight set) and the baseline approach

	Edge Count	Depth	CCC	GCS (%)	RFR (%)	GRC (%)	CAS (%)	SNI(%)
ISN (weight set 11)	6.09	0.56	1.26	6.65	91.13	6.60	85.29	2.31
ISN-AF (weight set 11)	6.09	0.51	1.15	6.90	94.22	7.23	94.70	2.89
Baseline	5.49	0.37	2.25	4.99	92.23	6.32	-	6.66

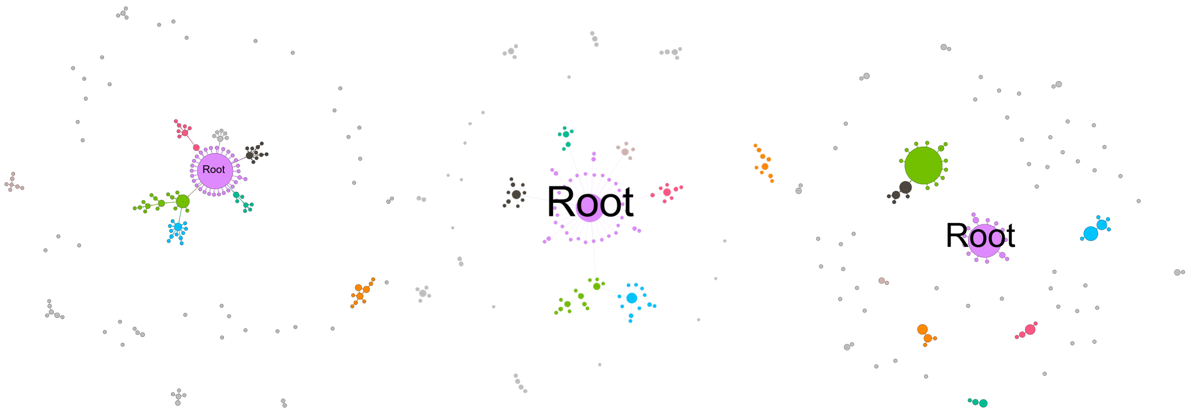
As a practical example, to highlight the difference between the baseline and the proposed approaches, we report four retweet cascade graphs with increasing node sizes. The cascade trees are implemented using Gephi version 0.9.2 and the node size represent the number of node's descendent. In other words, the bigger the node, the higher is its importance for tweet propagation.



(a) ISN **(b)** ISN-AF **(c)** Baseline

Figure 5. Retweet cascade graph comparison #1

Figure 5 reports the comparison of the retweet graphs generated by the three approaches with a total number of 67 nodes. While the baseline graph is characterized by a deeper tree with some sparse nodes, the proposed models' cascades (both ISN and ISN-AF) create a fully connected tree composed by only one level, since all the nodes are directly linked to the $root_{author}$. Considering a bigger cascade, Figure 6 compares the three approaches when the retweet number is 106: this result visually explains the considerations about the higher number of isolated connected component created by the baseline algorithm. In fact, from Figure 6c it is possible to observe the presence of numerous isolated connected components with nodes characterized by a valuable importance (node size) in terms of retweet propagation. It seems that the *root* has a minimum impact on the message propagation, while the other nodes (the big green and the big blue nodes) casually found the tweet and spread it. In contrast, the cascades in Figure 6a and Figure 6b propose a different retweet path, giving more



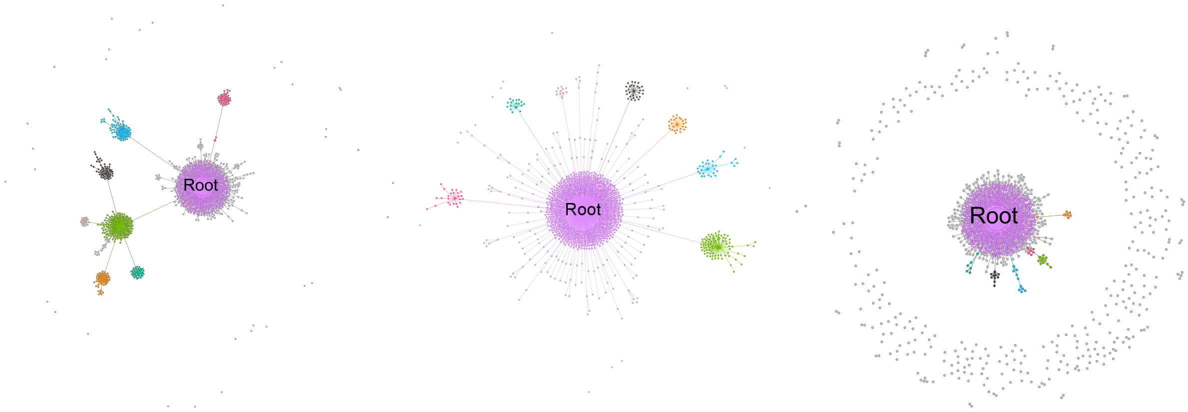
(a) ISN **(b)** ISN-AF **(c)** Baseline

Figure 6. Retweet cascade graph comparison #2

importance to the *root* and reducing also the number of single isolated nodes. In these two initial examples (Figure 5 and Figure 6) the proposed methods are showing similar or identical behavior, which are substantially different from the baseline.

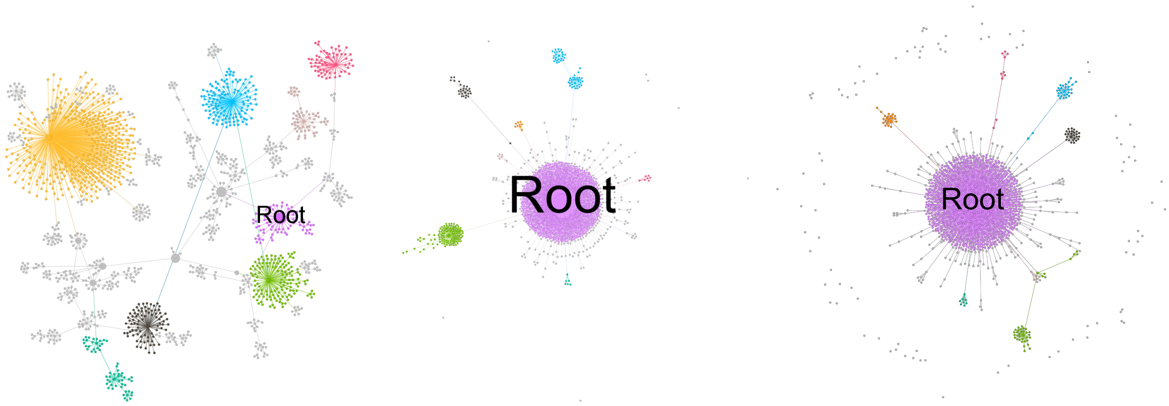
The other two examples (Figure 7 and Figure 8) report cascades where ISN and ISN-AF produce different propagation paths. In this case, the retweet number is 1,127 and 1,180, respectively. For the baseline approach (Figure 7c, Figure 8c) the node sizes, describing the nodes influence in the graph, favor the *root*. Conversely, the proposed ISN model shows propagation paths that highlight the importance of non-root nodes in retweet propagation. For instance Figure 7a shows the presence of important nodes (the big green and blue ones) that helped the propagation. These nodes are also present in the ISN-AF model (Figure 7b), but with a minor impact.

A more visible difference between ISN and ISN-AF is apparent from Figure 8b. The number of nodes connected to the root according to the ISN-AF approach are 621, of which 538 links are derived by IS interactions and 21 by friend network analysis. Differently, Figure 8a reports the ISN-generated graph, which has 1146 nodes linked by timelines information, thus, connected adopting the IS metric. Yet, the two graphs clearly diverge as, for the ISN model, a central role is covered by non-root accounts (big gray dots) that infected specific communities (the orange, blue and green components), whereas the ISN-AF cascade is dominated by the central role of the *root*, similar to the baseline graph (Figure 8c). However, in terms of IS metric the result is close, as the average CAS related to the 1146 edges derived by the ISN model is equal to 38.94%, while for the 538 edges related to the ISN-AF approach the CAS is 39.99%.



(a) ISN **(b)** ISN-AF **(c)** Baseline

Figure 7. Retweet cascade graph comparison #3



(a) ISN **(b)** ISN-AF **(c)** Baseline

Figure 8. Retweet cascade graph comparison #4

9. Conclusions

In this paper we have proposed two novel approaches to derive retweet cascade graphs, both based on the concept of interaction strength (IS). Differently from state-of-the-art approaches that are mainly based on users' social networks, the IS metric aims to enable a realistic estimation of Twitter information propagation considering the trust relationships among users, and giving more importance to frequent and strong interactions instead of considering the simple following relation. The two proposed approaches are the interaction strength-based network (ISN) and the interaction strength-based network with author's followers evaluation (ISN-AF). The ISN approach consists in two steps: the first phase aims to discover interactions among nodes involved in the retweet cascade, identifying links able to maximize the cascade average strength (CAS); then, for the remaining nodes the state-of-the-art approach based on nodes' social network is adopted. The ISN-AF approach also has a preliminary analysis of $root_{author}$ followers, which are used to generate the first level fo the cascade graph. This leads to an even more realistic graph, but also has some limitations.

The analysis and, in particular, the comparison with the baseline method show the differences between the proposed approaches, and their ability to highlight node importance in the retweet information propagation path. Notably, ISN and ISN-AF are able to reduce the impact of sparse nodes, diminishing the isolated connected components, and thus giving a more complete view of the retweet propagation phenomena. Hence, the study of the different propagation paths derived by ISN and ISN-AF (Figure 8) can offer a starting point for different future application such as social bot detection, inauthentic coordinated behaviours, etc. [31]. Moreover, differently from existing literature models, the proposed ISN and ISN-AF cascades can be evaluated through the CAS value, which gives to the analyst a proxy of the goodness of the obtained graph.

Notwithstanding the high number of experiments conducted, further analysis may be performed in order to optimize the interaction weights starting from the findings achieved in this paper. Other aspects that we intend to explore in future work are the tweet contents and the sentiment included in quotes and replies. In conclusion, despite the absence of ground truth information represents a limitation of the presented results, we believe that the proposed approaches represent a promising alternative solution to the problem of Twitter retweet cascade graph construction, which better highlights nodes importance during information flows and that overcomes the limitations of models based only on social network information.

Funding: "This research was partially funded by the EU project SoBigData-PlusPlus (grant no. 871042)"

References

1. Avvenuti, M.; Cresci, S.; Marchetti, A.; Meletti, C.; Tesconi, M. Predictability or early warning: using social media in modern emergency response. *IEEE Internet Computing* **2016**, *20*, 4–6.
2. Zola, P.; Ragno, C.; Cortez, P. A Google Trends spatial clustering approach for a worldwide Twitter user geolocation. *Information Processing & Management* **2020**, *57*, 102312.
3. Attia, A.M.; Aziz, N.; Friedman, B.; Elhusseiny, M.F. Commentary: The impact of social networking tools on political change in Egypt's "Revolution 2.0". *Electron. Commer. Res. Appl.* **2011**, *10*, 369–374. doi:10.1016/j.elerap.2011.05.003.
4. Zola, P.; Cortez, P.; Brentari, E. Twitter alloy steel disambiguation and user relevance via one-class and two-class news titles classifiers. *Neural Computing and Applications* **2020**.
5. Myers, S.A.; Sharma, A.; Gupta, P.; Lin, J.J. Information network or social network?: the structure of the twitter follow graph. 23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume. ACM, 2014, pp. 493–498. doi:10.1145/2567948.2576939.
6. Kwak, H.; Lee, C.; Park, H.; Moon, S.B. What is Twitter, a social network or a news media? Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010. ACM, 2010, pp. 591–600. doi:10.1145/1772690.1772751.
7. Comarela, G.; Crovella, M.; Almeida, V.A.F.; Benevenuto, F. Understanding factors that affect response rates in twitter. 23rd ACM Conference on Hypertext and Social Media, HT '12, Milwaukee, WI, USA, June 25-28, 2012. ACM, 2012, pp. 123–132. doi:10.1145/2309996.2310017.
8. Taxidou, I.; Fischer, P.M. Online analysis of information diffusion in twitter. 23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume. ACM, 2014, pp. 1313–1318. doi:10.1145/2567948.2580050.
9. Szabó, G.; Huberman, B.A. Predicting the popularity of online content. *Commun. ACM* **2010**, *53*, 80–88. doi:10.1145/1787234.1787254.
10. Yang, J.; Counts, S. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter. Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010. The AAAI Press, 2010.
11. Cogan, P.; Andrews, M.; Bradonjic, M.; Kennedy, W.S.; Sala, A.; Tucci, G. Reconstruction and analysis of twitter conversation graphs. Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research, 2012, pp. 25–31.
12. Yang, C.; Harkreader, R.C.; Zhang, J.; Shin, S.; Gu, G. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012. ACM, 2012, pp. 71–80. doi:10.1145/2187836.2187847.
13. Cazabet, R.; Pervin, N.; Toriumi, F.; Takeda, H. Information Diffusion on Twitter: Everyone Has Its Chance, But All Chances Are Not Equal. Ninth International Conference on Signal-Image Technology & Internet-Based Systems, SITIS 2013, Kyoto, Japan, December 2-5, 2013. IEEE Computer Society, 2013, pp. 483–490. doi:10.1109/SITIS.2013.84.
14. Zaman, T.; Fox, E.B.; Bradlow, E.T. A Bayesian Approach for Predicting the Popularity of Tweets. *CoRR* **2013**, *abs/1304.6777*, [1304.6777].
15. Pramanik, S.; Saha, A.; Mukherjee, P.; Patni, A.; Dan, S.; Mitra, B. Modelling Retweet Dynamics using Hawkes Process-a temporal approach **2015**.
16. Yu, L.; Cui, P.; Wang, F.; Song, C.; Yang, S. From Micro to Macro: Uncovering and Predicting Information Cascading Process with Behavioral Dynamics. 2015 IEEE International Conference on Data Mining, ICDM 2015, Atlantic City, NJ, USA, November 14-17, 2015. IEEE Computer Society, 2015, pp. 559–568. doi:10.1109/ICDM.2015.79.
17. Zhao, Q.; Erdogdu, M.A.; He, H.Y.; Rajaraman, A.; Leskovec, J. SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015. ACM, 2015, pp. 1513–1522. doi:10.1145/2783258.2783401.

18. Gao, J.; Shen, H.; Liu, S.; Cheng, X. Modeling and Predicting Retweeting Dynamics via a Mixture Process. Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume. ACM, 2016, pp. 33–34. doi:10.1145/2872518.2889389.
19. Kobayashi, R.; Lambiotte, R. TiDeH: Time-Dependent Hawkes Process for Predicting Retweet Dynamics. Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016. AAAI Press, 2016, pp. 191–200.
20. Rodrigues, T.; Cunha, T.D.S.; Ienco, D.; Poncelet, P.; Soares, C. RetweetPatterns: Detection of Spatio-Temporal Patterns of Retweets. New Advances in Information Systems and Technologies - Volume 1 [WorldCIST'16, Recife, Pernambuco, Brazil, March 22-24, 2016]. Springer, 2016, Vol. 444, *Advances in Intelligent Systems and Computing*, pp. 879–888. doi:10.1007/978-3-319-31232-3_83.
21. Cao, Q.; Shen, H.; Cen, K.; Ouyang, W.; Cheng, X. DeepHawkes: Bridging the Gap between Prediction and Understanding of Information Cascades. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017. ACM, 2017, pp. 1149–1158. doi:10.1145/3132847.3132973.
22. Zhou, Y.; Zhang, B.; Sun, X.; Zheng, Q.; Liu, T. Analyzing and modeling dynamics of information diffusion in microblogging social network. *J. Netw. Comput. Appl.* **2017**, *86*, 92–102. doi:10.1016/j.jnca.2016.09.011.
23. Stai, E.; Milaiou, E.; Karyotis, V.; Papavassiliou, S. Temporal Dynamics of Information Diffusion in Twitter: Modeling and Experimentation. *IEEE Trans. Comput. Social Systems* **2018**, *5*, 256–264. doi:10.1109/TCSS.2017.2784184.
24. Bhowmick, A.K.; Gueuning, M.; Delvenne, J.; Lambiotte, R.; Mitra, B. Temporal Sequence of Retweets Help to Detect Influential Nodes in Social Networks. *IEEE Trans. Comput. Social Systems* **2019**, *6*, 441–455. doi:10.1109/TCSS.2019.2907553.
25. Chen, G.; Kong, Q.; Xu, N.; Mao, W. NPP: A neural popularity prediction model for social media content. *Neurocomputing* **2019**, *333*, 221–230. doi:10.1016/j.neucom.2018.12.039.
26. Wu, B.; Cheng, W.; Zhang, Y.; Cao, J.; Li, J.; Mei, T. Unlocking Author Power: On the Exploitation of Auxiliary Author-Retweeter Relations for Predicting Key Retweeters. *IEEE Trans. Knowl. Data Eng.* **2020**, *32*, 547–559. doi:10.1109/TKDE.2018.2889664.
27. Arnaboldi, V.; Gala, M.L.; Passarella, A.; Conti, M. Information diffusion in distributed OSN: The impact of trusted relationships. *Peer-to-Peer Networking and Applications* **2016**, *9*, 1195–1208. doi:10.1007/s12083-015-0395-2.
28. Arnaboldi, V.; Passarella, A.; Tesconi, M.; Gazzè, D. Towards a Characterization of Egocentric Networks in Online Social Networks. On the Move to Meaningful Internet Systems: OTM 2011 Workshops - Confederated International Workshops and Posters: EI2N+NSF ICE, ICSP+INBAST, ISDE, ORM, OTMA, SWWS+MONET+SeDeS, and VADER 2011, Hersonissos, Crete, Greece, October 17-21, 2011. Proceedings. Springer, 2011, Vol. 7046, *Lecture Notes in Computer Science*, pp. 524–533. doi:10.1007/978-3-642-25126-9_64.
29. Nies, T.D.; Taxidou, I.; Dimou, A.; Verborgh, R.; Fischer, P.M.; Mannens, E.; de Walle, R.V. Towards Multi-level Provenance Reconstruction of Information Diffusion on Social Media. Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015. ACM, 2015, pp. 1823–1826. doi:10.1145/2806416.2806642.
30. Mones, E.; Vicsek, L.; Vicsek, T. Hierarchy measure for complex networks. *PloS one* **2012**, *7*.
31. Mazza, M.; Cresci, S.; Avvenuti, M.; Quattrociocchi, W.; Tesconi, M. RTbust: Exploiting Temporal Patterns for Botnet Detection on Twitter. Proceedings of the 11th ACM Conference on Web Science, WebSci 2019, Boston, MA, USA, June 30 - July 03, 2019. ACM, 2019, pp. 183–192. doi:10.1145/3292522.3326015.