A comparative study of social interactions and their combination on social media.

Io Taxidou
University of Freiburg
taxidou@informatik.
uni-freiburg.de

Peter M. Fischer
University of Augsburg
peter.fischer@informatik.
uni-augsburg.de

Martin Zablocki
Trivadis
martin.zablocki@
trivadis.com

ABSTRACT

Information diffusion on social media studies how information propagates from user to user considering temporal, spatial, social and structural aspects. The baseline of such analyses is diffusion graphs, i.e. *time-dependent graphs*, that show who is influenced by whom to propagate a piece of information. In this paper we classify, compare and complement user interactions and influence on social media from our own previous works and state-of-the-art. In contrast to previous work that studied diffusion along a single means of transfer, we investigate and evaluate diffusion graphs considering different types of interactions at the same time. We show that such analysis yields more complex structures and a more complete picture of diffusion.

PVLDB Reference Format:
Io Taxidou, Peter M. Fischer, Martin Zablocki. A comparative study of social interactions and their combination on social media.. *PVLDB*, 11 (9): xxxx-yyyy, 2018.
DOI: https://doi.org/TBD

1. INTRODUCTION

Social media, micro-messaging services, or sharing sites (e.g., Facebook, Twitter, or Instagram) provide the means of interactions among people in which they create, share, and exchange information and ideas in virtual communities and networks. Many real-life situations, such as elections [24] or natural disasters [30] are reflected on social media. In turn, social media shape these situations by forming opinions, strengthening trends or by spreading news on emerging situations faster than conventional media.

Those phenomena are studied by information diffusion, i.e., tracing, understanding and predicting how a piece of information is spreading. Such analysis provides valuable insights on who is propagating certain information and who is influencing others. In the literature, information diffusion is modeled as *information cascades*: information cascades are time-dependent graphs that show who was influenced by whom in order to propagate a piece of information. Time-dependent graphs bear certain restrictions, since time plays a key role in both modeling and evaluating them. In particular, those graphs have a strong temporal component in two dimensions. In the macro level, influence fades out very quickly and information is soon outdated, which should be considered when analyzing information diffusion. In the micro-level by looking at individual users, the point in time they will get involved determines their role in the information diffusion eco-system. For example, celebrities or information sources tend to pick up information very quickly and influence many others [33].

There is plenty of literature that studies information diffusion ranging from building models that approximate those processes [11, 15, 21], inferring diffusion paths [9], predictions for information diffusion like size [20], speed [36], burst [19] etc., influential identification [5] and event/ trend detection [37, 4, 23]. However, in most cases one type of diffusion is being considered, like retweets in Twitter, re-shares in Facebook, propagation of memes, etc. In previous work, we have shown that there exist multiple types of interactions in social media and by focusing in a single type, we are missing the big picture and the full potential of information diffusion [34].

The first contribution of this paper is to *compare and classify different types of interactions and influence*. We consider both *explicit interactions*, based on social media providers provenance information i.e., mechanisms like retweet in Twitter or reshare in Facebook and *implicit interactions*, based on latent influence indicators, for example users conventions of crediting their sources, social connections or content similarity. In order to compute user interactions and influence we need to identify the *provenance* of messages, i.e. the sources, the intermediate steps and any modifications that a piece of information has undergone on the way. The second contribution is a newly presented method for *loosely computing influence*, which is attributed to the impact of the content and not to individual users. This method is based on *hashtag propagation*, i.e. tagged phrases that propagate from user to user. As a third contribution, we take a step further and analyze the *combination of different types of user interactions*, which has been only studied in isolation in state-of-the-art. Our results show that when we consider combined interactions, we observe more complex structures that were not revealed by investigating them in isolation. Those results should be taken into account, when modeling the processes of information diffusion or analyzing information cascades.

To summarize, the contributions of this paper are the following:

- We comparatively study and analyze user interactions in social media. In particular, we categorize them and we discuss them under the lenses of different dimensions (for example: source, previous step, computational assumption, etc)

- We provide a new method for loosely computing influence by tracing hashtags, which constitutes a key aspect for the context of the message.

Proceedings of the VLDB Endowment, Vol. 11, No. 9
Copyright 2018 VLDB Endowment 2150-8097/18/5.
DOI: https://doi.org/TBD

- We evaluate combined interactions and we compare them with single-type interaction cascades.

The remainder of the paper is structured as follows: we describe our methodology in section 2 where we explain the different methods of interactions and how they were combined. Section 3 presents the results of our analysis while we provide related work in section 4. Section 5 concludes the paper.

2. METHODOLOGY

In this section we describe our methodology for computing information cascades and how we combine them. We leverage several of our previously developed methods which we compare, extend and put into context in this paper.

For information cascade reconstruction we are based on the following assumptions:

- Information cascades are modeled as directed graphs where nodes represent users and edges show influence to further propagate a piece of information. The edge direction demonstrates influence flow starting from the ”influencer” and pointing to the ”influencee”.

- For every information cascade there exists at least one root, which is the starter of information diffusion or the first user to post particular information.

- We assume the existence of an underlying social graph for some of our methods. For others, the existence of a social connection strengthens our inference.

- We allow the existence of multiple influencers as we believe is close to reality: online users are consuming information from multiple sources and are not influenced by a single source most of the times.

Next, we categorize our methods according to *implicit* and *explicit* interactions and we present our hypothesis for reconstructing such interactions. In order to do that, we need to compute who is influenced by whom to propagate some information, in other words the *provenance* of information.

Some hypotheses are competing (alternative hypotheses) and they cannot be tested at the same time. Alternative hypotheses can be sorted, ranked, weighted according to their plausibility and use case. Other hypotheses are complementary, which means that they can be applied on top of other existing hypotheses. They aim at providing more fine-grained provenance information, while strengthening the already computed provenance.

As a use case and for the sake of providing examples, we focus on Twitter for both explaining and evaluating our methods. However, those methods can be applied to other social media, since similar bevariours and interaction patterns have been observed to other social platforms. Next, we present the methods that we used to collect data, which have an impact on which type of provenance information is provided and which needs to be inferred.

2.1 Dataset Collection

Performing an 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.

2.1.1 *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 geographical coordinates. 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, as a result we are missing messages (and retweets). Furthermore, Twitter provides only limited means to retrieve messages after their occurrence.

These limitations have another consequence: we cannot observe all possible cascades, but need to settle for specific subsets before we start to record. For that, 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 and we leave for future work. Instead, 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 or US elections), 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.

2.1.2 *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. We favor those users that are retweeted most and/or have the most followers, in order to capture possible popular users that exert influence on others [7]. 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. We retrieve the friend information when needed explicitly for some variants of our algorithm.

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

2.2 Explicit Interactions

Explicit interactions refer to those cases where users are attributing credit to their influences through social media operators, like retweet in Twitter. For explicit interactions, we observe two main cases:

- *Direct linkage based*, like replies and quotes in Twitter where the previous step[2] is provided.
- *Source based*, like retweets in Twitter where the source[3] is provided.

Figure 1 shows these types of interactions. *Single-step edge* notifies that a message is derived through one step from another, while *any-step edge* means that a message is derived through one or more steps from another.

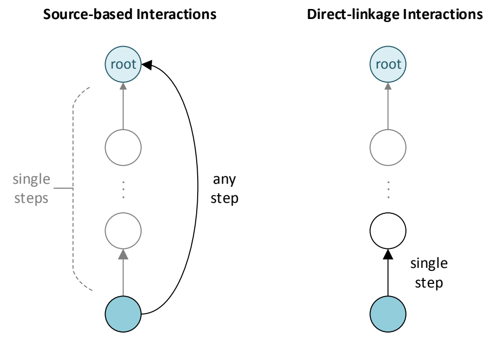


Figure 1: Explicit Interactions where nodes depict messages. The information that is given by social media providers is highlighted in black colour. The information that needs to be inferred for fully tracing the diffusion paths is depicted in grey. For the source based interactions on the left, the reference/edge to the root is called *any-step* because the root is being reached through one or more predecessors. The *single-step* edges that lead to the root need to be inferred. For the direct linkage based on the right, a *single step* edge links to the previous message, which is given. The root has to be inferred.

Orthogonal aspects to this categorization are:

- the number of *possible roots*: single vs multiple roots
- the number of *influence edges* that are generated: single vs multiple edges

For the *direct linkage interactions*, since we know the previous step, we have very low uncertainty for the influence edges. However, the source is not provided which has to be found by following back the influence paths. We present such methods in [31] where we describe how to reconstruct replies and quotes in Twitter

For the source-based interactions, where the source is provided, we need to infer the influence paths that lead to the source. We also know that those messages belong together, because the source is being referenced.

In order to make the inference, we make the hypothesis that

H1: Influence flows through social connections.

We rely on this hypothesis due to the nature of social media: users are exposed to the content shared by their social connections. As a result, they are most likely to be influenced by them. The same findings also are supported in the literature [14, 8]. This hypothesis applies to source-based interactions, and in general to cases where the previous influencer is not provided, like in the case of implicit interactions. These influence edges have medium to low uncertainty, since users might also get influenced from the public timeline, without any obvious connection [26]. Note here, that more than one activated social connections lead to multiple diffusion paths: a large number of alternative diffusion paths, raises respectively the uncertainty of the reconstructed influence edges (and as a result the influence paths). We present source based methods and in particular retweet reconstruction methods in [32, 10].

In terms of *number of roots* and *number of generated edges*, in cases where the source is provided, we observe a single source and possibly multiple influence edges.

For direct linkage, the root is not provided, but since for every message one influencer is embedded, this will result in a single root. When the direct linkage is provided, we might end up having multiple roots, since there is no explicit single root. The number of influence edges is also one, because it is embedded in every message. Note here that in some cases a message might carry two different direct linkage interactions (e.g. reply and quote in Twitter at the same time) which results in two embedded edges. In such a case, we encounter two roots.

2.3 Implicit Interactions

In this section we provide the background for implicit interaction and we consider two cases: in section 2.3.1 we present methods for exact implicit provenance computation from user to user. In section 2.3.2 we provide a new method for loosely attributing influence as a combination of previous adopters and virality of trends.

2.3.1 Exact provenance computation

For implicit interactions, we assume that users are influenced by an unidentified source (other users or external sources) but do not express it with the mechanisms of social media. As a result, there might exist some "hidden" provenance with regard to their messages, which we are trying to reveal.

Our basic hypothesis is that:

H2-I: If two messages are highly similar, there is a high probability that they share some provenance.

which leads us to the complementary following hypothesis:

H2-II: The higher the similarity between two messages, the higher the probability that they share some provenance.

Hypotheses H2-I and H2-II: are relevant for cases where no provenance information is provided (in contrast to Section 2.2) and we rely solely on the content of the messages. We identify those connections by a clustering algorithm (SimClus [3]) that clusters content-wise similar messages. The algorithm produces possibly overlapping clusters sharing similar context. As a result, a lightweight topic detection is a by-product of the clustering but out of scope for this work.

Within those clusters we identify latent influence and provenance. By using hypothesis H2-II we assume that highly similar messages must share some provenance and we connect every message with its most similar that was written in the past (*single-step provenance*). We also assume that all the messages within each cluster are derived through one or more steps from the oldest message in each cluster (*any-step provenance*). Figure 2 depicts the provenance generated in implicit interactions. This way, we reduce the num-

[2] Direct predecessor

[3] Origin of diffusion. Here we make the distinction between the *source* and the *root* of diffusion: by source we mean the true origin of diffusion, in an ideal case where we posses all the relevant data. In cases of missing data, the *observed* origins of diffusion are called roots. In the ideal scenario, the source and the root are aligned.

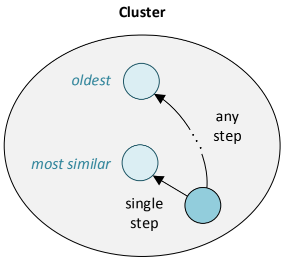


Figure 2: Implicit Interactions in one cluster. Since a cluster refers to a topic and the messages within the cluster are more similar with the messages across other clusters, we assume that all the messages are derived through one or more steps from the oldest messages in the cluster (*any-step*). Additionally, Every messages is assumed to be derived through *one-step* from its most similar oldest one.

ber of possible roots (single root: oldest message in the cluster) and the number of possible influence edges (singe edge: most similar previous message). However, if two clusters are overlapping the number of roots and influence edges is doubled for the messages in the overlap (one root and one edge for each cluster).

This hypothesis still carries high uncertainty, since content similarity does not always imply provenance and influence. In order to further decrease the uncertainty of the previous methods we form the additional hypothesis:

H3: Online users use their own conventions to express influence and provenance.

Hypothesis H3 can be applied to any message in order to unravel latent influence or strengthen the already computed influence. With respect to H2 and H3 we categorize implicit interactions accordingly:

- *Content similarity* based: similar messages that share some provenance.
- *Additional indicators* based: applied on top of the similarity (or as standalone) which lowers the uncertainty of the reconstructed provenance.

Note here that in both cases, we rely on inference, since no provenance information is provided from social media. For the orthogonal categorization on the *number of roots* and *the number of influence edges*, in theory we have multiple roots and influence edges, since no information is provided and we rely on inference. However our methods presented in [35] consider the grouping of messages that belong together under a single root and we compute the most plausible previous linkage edge. Our methods allow overlapping clusters and in this case, we observe more than one roots (depending on the overlap).

For the *content based similarity* method, the uncertainty is high because our methods are based solely on the similarity among messages. In cases of external (out of social media) influence (e.g. major events) many influence edges are created, which do not necessarily imply influence from other users. In this work, we quantify and model external influence but we plan to properly compute it in future work.

In the *additional indicators* based method, we leverage particular indicators to decrease the high uncertainty of the previously reconstructed provenance. For example, user mentions or the existence of social connections strengthen our hypotheses of the inferred influence. For more details about additional indicators we refer the reader to [35]. The generated provenance edges bear medium to low uncertainty according to the strength of each indicator.

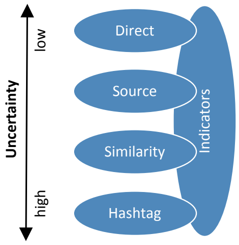


Figure 3: Our methods ranked according to uncertainty of inference.

In Figure 3 we provide the methods ranked according to the uncertainty of results. Additional indicators overlap with the rest, because they borrow methods and assumptions from the other three categories, which leads in varying uncertainty according to the method used. Note that our aim is not to build an uncertainty model or quantify it; we rather desire to understand what is the strength of those influence edges and which additional hypothesis might lower the uncertainty involved.

2.3.2 Loose provenance computation

In this section we present a method for loosely computing provenance. i.e. not attributed to particular user but to the influence inflicted from particular information in the past. A complementary means of transfer that we present in this paper is *hashtag reconstruction*. Hashtags include annotated words or combinations of words that characterize the message and refer to a particular situation (event, meme, etc). Hashtags are meanwhile adopted by most social media platforms including Twitter, Facebook, Instagram, Pinterest and Linkedin.

We categorize hashtags as implicit means because we consider them as user conventions to annotate keywords. Although hashtags might be considered as explicit interactions because the words followed by the hashtag symbol *"#"* are traced explicitly by social media, there is no explicit credit attribution that refers to specific provenance. The purpose of a hashtag is to provide visibility and contribute to trends. From a computational point of view, hashtag reconstruction belongs to the family of similarity based methods, since hashtags are textual information. We decided to explicitly reconstruct hashtags as they are considered to be representative of a user's message. This way, we account for cases where the similarity is too weak to create provenance with the similarity based method.

Here we consider a more weak notion of provenance: instead of considering user to user influence, we assume that the aggregate popularity of a certain hashtag drives its further propagation.

We are based on the following hypothesis to reconstruct hashtags:

H4: The greater the popularity of a hashtag, the most likely is that users are adopting it. As a result, we consider the aggregate hashtag influence and assume that a user is influenced by all previous hashtag users.

We consider the users who post a hashtag are influencing all subsequent users who also use the same hashtag. In this sense, we assume that the wider the use of the hashtag, the higher the possibilities that is adopted by others. Those edges carry high uncertainty, as seen in Figure 3, while influence is attributed not to the individual users but to the impact or visibility of the hashtag. This method

Table 1: Overview and Properties of our Methods

	Source	Direct linkage	Similarity	Additional Indicators	Hashtag
General Assumption	credit to influencer	credit to influencer	Influence without any credit	user conventions, indicators	hashtag visibility
Source (#)	1, provided	≥1, not provided	≥1, not provided	≥1, not provided	≥1, not provided
Previous step (#)	≥1, not provided	≥1, provided	≥1, not provided	≥1, not provided	≥1, not provided
Computational Assumption	social connections	-	content similarity	mention, social connection, interaction with explicit means	all previous hashtag users
Edges generated (#)	≥1 (# of connections activated)	≥1	2 (oldest/ similar in each cluster)	≥1	≥1
Uncertainty	medium-low	low	high	medium-low	high
Grouping (#)	retweet cascade (1)	reply/ quote cascade (1 or 2)	topic clusters (≥1)	topic clusters (≥1)	hashtag cascades (≥1)

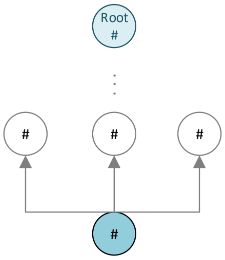


Figure 4: Hashtag Reconstruction: influence is attributed to all previous users, in other words to the popularity of the hashtag

shares similarities with statistical models like the Linear Threshold model [15], where a user becomes influenced if the total weight of its incoming neighbors is higher than the threshold.

Next, we provide some details how we computed the provenance of hashtags. Coming back to our scenario of collecting datasets, the search keywords we use for crawling the data very often constitute hashtags or parts of them. When we reconstruct hashtags, we remove those that were used for crawling in order to avoid creating unnecessary connections, since they are contained in the majority of the messages. This way we focus on diverse hashtags that differentiate the fine grained topics belonging to the wider event or situation that we are crawling (like Olympics or Elections). Hashtag reconstruction is depicted in Figure 4 and shows the influence flowing from all previous hashtag users.

Table 1 summarizes the characteristics of hashtag reconstruction. Concerning the *number of roots* and the *number of generated edges* we observe multiple in both cases, which have to be inferred. We observe multiple roots where more than one hashtags are included in the reconstructed messages. For every hashtag, we observe a single root, which is the oldest message that used this hashtag. The number of influence edges depends on how many previous users have included the particular hashtag in their messages. The grouping of messages is the use of an individual hashtag, but in practice, users are mentioning more than one hashtags in their messages, and we observe tightly connected cascades of multiple hashtags.

It is worth to mention that we tested the hypothesis *H1* for hashtag reconstruction, and we used the social graph to unravel influence, i.e. provenance. We fail to confirm such hypothesis, since very few edges were attributed to social connections. However, social connections and in general any *influence indicator* can be applied additionally to strengthen the reconstructed provenance. We summarize our findings from comparing different types of implicit and explicit interactions in Table 1.

2.4 Combined Interactions Reconstruction

In our previous works and related work, those diffusion graphs (information cascades) have been investigated in isolation. When we evaluate those information cascades in combination, we observe provenance edges among the disconnected single type cascades. We ignore the labels, i.e. different semantics that those interactions bear and we consider simple edges.

In particular, the observe the following types of possible edges that connect isolated cascades (considering Twitter).

- *Explicit - Explicit*: A reply can be retweeted (and vice versa) and in this case a reply cascade gets connected with a retweet cascade (in particular with the root of the latter)
- *Explicit - Implicit*: A retweet root cascade or a reply can be connected with other messages through implicit means (e.g. similarity or use of hashtag)
- *Implicit - Implicit*: A similarity based cascade can be connected with a hashtag cascade with the use of similar hashtags in both cases. This happens when the similarity is too weak to create provenance but the use of common hashtags strengthen the assumption that particular messages share provenance.

By connecting cascades that consider a single type of interaction, we cannot refer to individual cascades any more but to a forest whose components include cascades of possibly combined interactions. This is a multi-graph, which means that we observe one or more labels over its edges. When computing different metrics in Section 3 we consider a simple graph and we do not treat edges differently. We output a single, unlabeled edge among two users, if any type of edge exists in the multi-graph. In other words, we project the multi-graph to a regular graph using existential semantics. However, according to the use case, one can consider only certain types of edges. For example if uncertainty is a key concern, only edges with lower uncertainty (or of a particular type) can be considered.

3. EVALUATION

In this section we evaluate some basic graph metrics in order to provide a first impression about the range of different interactions.

The dataset we used is taken from Twitter and it was recorded during the ISWC conference in 2015. The dataset contains 3909 messages, consisting of 2068 retweets, 198 quotes, and 93 replies.

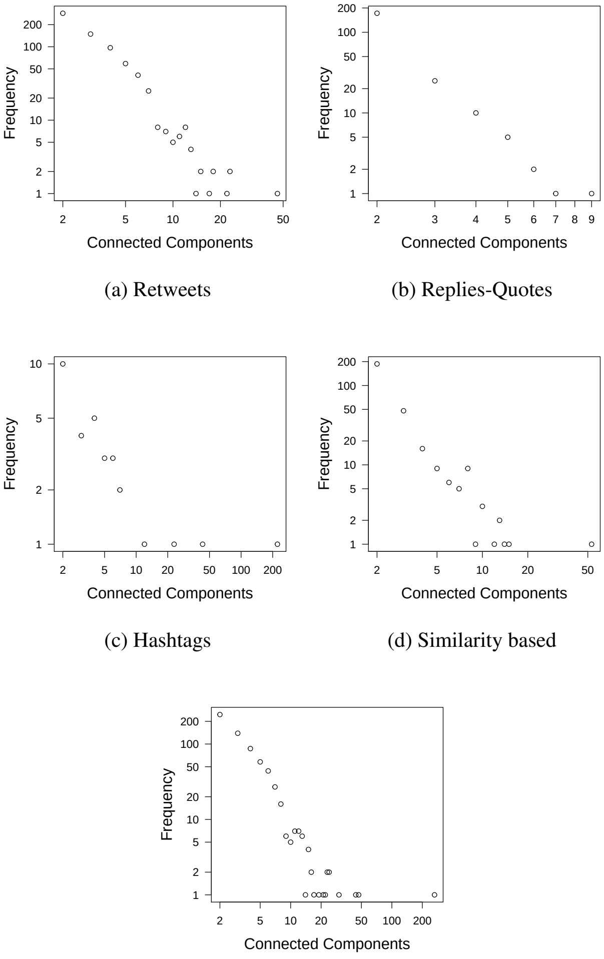
We used our methods outlined in the previous section (more details in [32, 31, 35]) in order to reconstruct explicit and implicit diffusion information diffusion cascades correspondingly. For retweets, as discussed in Section 2.2, the root of diffusion is provided but the intermediate forwaders not. We leveraged the assumption that influence flows through social connections (H1). In particular, we used the social graph (follower lists) crawled Twitter to identify who is connected with whom and compute influence[32]. For replies, since the previous step of diffusion is provided, we iteratively followed the paths that lead to the conversation root. In order to get additional replies, not included in our dataset, we crawled the involved users' timelines as described in [31]. For implicit diffusion reconstruction, we used SimClus as a clustering algorithm first presented in [3] and modified by us in [35] in order to cater for the hypothesis *H2-II*. In particular the goal is to maximize the similarity within each cluster in order to find more fine-grained connections. SimClus clusters similar messages using a lower bound of similarity, 0,4 in this evaluation, which we empirically selected.

For hashtag reconstruction we leverage the hypothesis *H3* and method presented in Section 2.3.2 using a more "loose" definition of influence. For that, we assume that all previous users that used a particular hashtag have contributed to subsequent users being influenced. When combining all types of interactions we observe the edge types described in Section 2.4 which connect the single type interaction cascades. Note that these edges exist in the isolated cascades, but since one type of interaction was considered we were ignoring those additional edges. For example, when evaluating retweet cascades [32], we considered only messages that are retweets and ignored the rest.

After we have constructed information cascades by computing provenance and influence, both as single interactions and combined we evaluate some basic social network analysis metrics. First, we present the cascade size distribution with respect to different types of diffusion and their combination. For the combined interactions, different single type cascades are interconnected and we refer to them as *connected components*. For consistency, we use the terms *connected components*, for single type and combined interactions. Note here that for single-type interaction cascades those connected components correspond to distinct information cascades.

Figure 5 shows the size distribution of connected components. We observe that the largest connected components of single interactions have not been connected when combined. In particular, the size of largest connected component in combined interactions has size of 264 (Figure 5e), while the size of the largest connected components for single interactions is 222 for hashtags (Figure 5c). Likewise, the amount of two sized connected components is massively decreased, given that the corresponding size for combined interactions is around 200, while for retweets, replies and similarity-based interactions we observe sizes around 200 for each of them. We also observed that the number of middle sized components (size 10-30) is also increasing for combined interactions.

This means that when combining interactions the smallest sized components are affected the most and not the largest. This is a surprising insight if we take into account the preferential attachment, i.e. the "rich is getting richer" effect that has been observed in social graphs (and in many other types graphs) [6]. In particular, high degree nodes attract more connections than low degree nodes. Similarly, high influential user tend to increase their influence faster than others. However, this theory does not apply here.



(a) Retweets

(b) Replies-Quotes

(c) Hashtags

(d) Similarity based

(e) Combined Interactions

Figure 5: Distribution of Component Sizes

Table 2 presents the diameter for the largest connected components for all types of interactions. The last row shows that the diameter for the combination of interactions is not increased as one would expect, but rather shrinks. This is partially explained due to the fact that the largest connected components are not merged, but also by the fact that alternative connections are created shortening the diffusion paths. As a result, we observe that information is faster transferred when considering multiple means of interactions.

Table 2: Diameter for different types of interactions

Interaction	**Diameter**
Retweets	8
Replies/Quotes	9
Hashtags	8
Similarity	5
Combined	8

Next, we look at the betweenness centrality which measures the fraction of shortest paths that pass through a node in Figure 6. Since the betweenness values are very small (75th percentile is 0, with exception the hashtag interactions which is 0,01), we are looking at high outliers. The number of high betweenness nodes is increasing slightly for the combined interactions, but still remains very close to the corresponding values for retweet cascades. This supports further the hypothesis that smaller cascades are becoming connected, as a result the betweenness centrality does not change drastically.

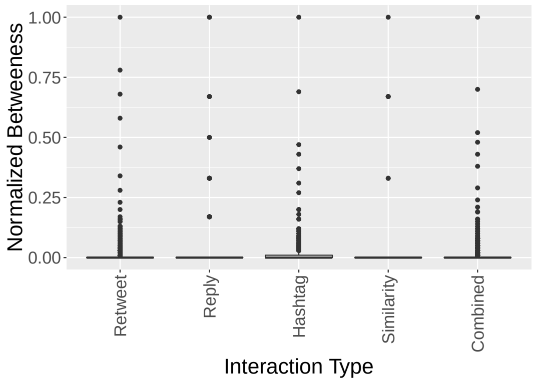


Figure 6: Betweeness Centrality

4. ADDITIONAL RELATED WORK

There is a lot of literature that studies information diffusion and information cascades. In particular a wide range of models have been developed considering multiple aspects of diffusion and making different assumptions. Modeling information diffusion has two main goals: from one hand to understand the underlying process and its evolution and from the other side to implement predictions based on such models. Many statistical models have been developed that simulate information diffusion processes, borrowed from epidemiology and the spread of diseases [11, 15, 16, 27] and constitute the baseline for more elaborate models, for example [1].

Here we are focusing on research on inferring information cascades aims in unraveling their underlying structure given a sequence of activation. The models developed here reconstruct the path taken by a piece of information. For example the continuous work of Manuel Gomez-Rodriguez [12, 28, 13] studies this problem focusing on different dimensions: the work in [12] reconstructs the social graph and cascades over which information propagate. The authors build a model that finds the spreading cascades by maximizing the likelihood of observed data. Every node can activate its neighbour independently based on some probability. Particularly, by observing many different cascades spreading from node to node, the edges of the underlying graph (and as a result cascades) can be inferred. The datasets used for evaluation include memes propagating over blogs and news websites. The results indicate a core-periphery structure of the underlying social graph and large influence from mass media is identified. The authors extend their model in [28] in order to account for the times and rates of transmission of individual edges rather than having a uniform probability for each edge. Such a problem has been addressed before [25] but with certain assumptions and heuristics, while the solution of [28] does not require parameter tuning.

In these two works [12, 28] the authors assume that the underlying social graph stays the same, which is not a correct assumption in reality. In [13] the authors account for the dynamics in terms of structure and timing of the underlying social graph. The underlying social graph is unraveled by observing the infection times of nodes. The authors analyse the dynamic information diffusion paths and concluded that they are more stable for general recurrent topics, while real-world events results in large changes over these pathways. The work in [22] takes a different approach in modeling the dynamics of the social and diffusion graphs: under the assumption that individuals tend to break old connections and connect to the their two hop friends, a link rewiring strategy is implemented. Simulations are under the SIR model and with the developed strategy show that information spreads faster and deeper.

While the previous methods were based on modelling, the work by Cogan et al. [9] studied user interactions on Twitter by reconstructing the conversational graphs of mentions, retweets, replies by means of observable edges. Their dataset is much smaller compared to what we target: it contains 33K retweets while the largest retweet cascade has size 170 retweets. A similar approach is taken by [17] to reconstruct retweet cascades from tweets that explicitly mention the source in their tweets (*@username*). That was the old convention of retweeting before the official means released. While the previous works inferred information cascades out of individual activations which were assumed to be complete, the work of [29] accounts for missing data. The authors build a model to estimate the overall properties of cascades even in cases up to 90% of data are missing. Instead of inferring the overall properties, the work in [38] infers missing nodes by incorporating temporal information. However, the cost of such inference is quite significant, depending on the size of the entire social graph.

5. CONCLUSION AND FUTURE WORK

In this paper we investigate different types of interactions and user influence on social media. In particular, we trace information diffusion through different means by identifying its provenance. We categorize and compare those different methods and we present a new method for hashtag reconstruction. We evaluate those methods in isolation and in combination, which is a contribution of this paper. Our results show that small to middle sized information cascades are mainly affected when evaluating combined interactions. This observation comes in contrast to the well established preferential attachment theory of social networks.

Here we presented a preliminary evaluation with combined interactions cascades and more research is needed into this directions to identify what are the implications of our preliminary findings.

For example: How information diffusion statistical models or link prediction models are affected? What is the impact in influence problems, i.e. influence maximization[18]? Could we better understand topics or event/trend detection tasks? How our findings will affect the recent reserach in evolution of ego-networks [2]?

For future work, we plan to apply a wider variety of social network analysis metrics to better understand the implications of considering more than one type of interaction. Given the multiple types of interactions and influence, the next step is to quantify *uncertainty* of influence computation. By doing this we will get important insights for influence on social media especially concerning the strength of such connections.

6. REFERENCES

[1] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In *Proceedings of the 2008 international conference on web search and data mining*, pages 207–218, 2008.

[2] L. M. Aiello and N. Barbieri. Evolution of ego-networks in social media with link recommendations. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 111–120. ACM, 2017.

[3] M. Al Hasan, S. Salem, and M. J. Zaki. Simclus: an effective algorithm for clustering with a lower bound on similarity. *Knowledge and information systems*, 28(3):665–685, 2011.

[4] F. Alvanaki, S. Michel, K. Ramamritham, and G. Weikum. See what's enblogue: real-time emergent topic identification in social media. In *Proceedings of the 15th International Conference on Extending Database Technology*, pages 336–347. ACM, 2012.

[5] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74, 2011.

[6] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[7] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10(10-17):30, 2010.

[8] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, pages 721–730. ACM, 2009.

[9] P. Cogan, M. Andrews, M. Bradonjic, W. S. Kennedy, A. Sala, and G. Tucci. Reconstruction and analysis of twitter conversation graphs. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, pages 25–31. ACM, 2012.

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

[11] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223, 2001.

[12] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1019–1028, 2010.

[13] M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf. Structure and dynamics of information pathways in online media. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 23–32, 2013.

[14] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM, 2010.

[15] M. Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.

[16] H. W. Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

[17] E. Kafeza, A. Kanavos, C. Makris, and P. Vikatos. Predicting information diffusion patterns in twitter. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 79–89. Springer, 2014.

[18] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, 2003.

[19] S. Kong, Q. Mei, L. Feng, F. Ye, and Z. Zhao. Predicting bursts and popularity of hashtags in real-time. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 927–930. ACM, 2014.

[20] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev. Prediction of retweet cascade size over time. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2335–2338, 2012.

[21] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. In *Proceedings of the 2007 SIAM international conference on data mining*, pages 551–556, 2007.

[22] C. Liu and Z.-K. Zhang. Information spreading on dynamic social networks. *Communications in Nonlinear Science and Numerical Simulation*, 19(4):896–904, 2014.

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

[24] P. T. Metaxas and E. Mustafaraj. Social media and the elections. *Science*, 338(6106):472–473, 2012.

[25] S. Myers and J. Leskovec. On the convexity of latent social network inference. In *Advances in Neural Information Processing Systems*, pages 1741–1749, 2010.

[26] S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–41, 2012.

[27] M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[28] M. G. Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal dynamics of diffusion networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 561–568, 2011.

[29] E. Sadikov, M. Medina, J. Leskovec, and H. Garcia-Molina. Correcting for missing data in information cascades. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 55–64, 2011.

[30] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[31] I. Taxidou, T. De Nies, and P. M. Fischer. Provenance of explicit and implicit interactions on social media with w3c prov-dm. *Lecture Notes in Computer Science, Springer LNCS/LNAI series*, 2017. Accepted, under publication procedure.

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

[33] I. Taxidou and P. M. Fischer. Structural aspects of user roles in information cascades. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1505–1509, 2017.

[34] I. Taxidou, P. M. Fischer, T. De Nies, E. Mannens, and R. Van de Walle. Information diffusion and provenance of interactions in twitter: is it only about retweets? In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 113–114. International World Wide Web Conferences Steering Committee, 2016.

[35] I. Taxidou, S. Lieber, P. M. Fischer, T. De Nies, and R. Verborgh. Web-scale provenance reconstruction of implicit information diffusion on social media. *Distributed and Parallel Databases*, pages 1–33, 2017.

[36] J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in twitter. *Proceedings of the 10th International AAAI Conference on Web and Social Media*, 10:355–358, 2010.

[37] X. Zhou and L. Chen. Event detection over twitter social media streams. *The VLDB Journal - The International Journal on Very Large Data Bases*, 23(3):381–400, 2014.

[38] B. Zong, Y. Wu, A. K. Singh, and X. Yan. Inferring the underlying structure of information cascades. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1218–1223. IEEE, 2012.