Realtime Analysis of Information Diffusion in Social Media

Io Taxidou
supervised by Peter Fischer
Web Science Group
CS Department, University of Freiburg, Germany
taxidou@informatik.uni-freiburg.de

ABSTRACT

The goal of this thesis is to investigate real-time analysis methods on social media with a focus on information diffusion. From a conceptual point of view, we are interested both in the structural, sociological and temporal aspects of information diffusion in social media with a twist on the real time factor of *what is happening right now*. From a technical side, the sheer size of current social media services (100's of millions of users) and the large amount of data produced by these users renders conventional approaches for these costly analyses impossible. For that, we need to go beyond the state-of-the-art infrastructure for data-intensive computation. Our high level goal is to investigate how information diffuses in real time on the underlying social network and the role of different users in the propagation process. We plan to implement these analyses with full and partially missing datasets and compare the cost and quality of both approaches.

Keywords

Social Media, Information Diffusion, Realtime Analysis

1. INTRODUCTION

Social media provides the means of interactions among people in which they create, share, and exchange information and ideas in virtual communities and networks. The advent of multiple platforms and the constant engagement of users on them has provided scientists with a treasure of information for analysis. For example, Facebook reported over a billion users and Twitter produces over a 400 million messages per day. Moreover, mobile technologies offer easy access everywhere and events are reported while happening [37]. Therefore, we can observe a huge amount of data being produced at incredibly rapid rates while information becomes faster outdated.

Up to now, social media analysis is implemented mainly offline and in relatively small datasets, with a few notable exceptions such as trend detection [4]. However, taking into account the rapid growth of social media which produces huge datasets with a strong temporal and structural aspect, new methods and systems for analysis are needed. We plan to work on interdisciplinary approaches including algorithms and methods from Social Network Analysis (SNA) while using techniques and systems that the Database community is developing. We strongly believe that the Social Network Analysis community could benefit from research in the field of Databases and Systems and address problems for both scientists and end users. A specific area to investgate is *information diffusion*, which studies the methods and mechanisms of how information propagates. Information diffusion has been a very active field of research especially after the advent of online social networks [27, 24, 47, 19, 26]. A recent survey [22] has brought the topic into visibility into the Database Community.

Social media platforms facilitate the observation and analysis of information diffusion due to the explicit network of users they maintain. Social media includes social relationships between users (e.g. friends, followers) over which multiple interactions take place. These interactions can be modeled as a graph the so called *information cascades*. Information cascade graphs provide a precise modeling of information spreading, where the edges reveal the relationship of "who was infected/influenced by whom". In Figure 1 an information cascade is depicted that unfolds over the the underlying social network. The infected nodes that propagate the message are coloured in red.

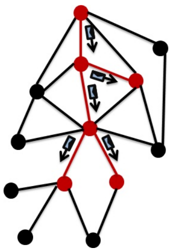

Figure 1: Information Cascade

We are focusing on the Twitter for the following reasons: Firstly, Twitter is a microblogging service that that resembles news media [25] and functions like a social sensor for events and trends. Secondly, Twitter's function of *retweeting*, which is forwarding another user's tweet message by giving credit to it, facilitates explicit information diffusion from one user to the other. Finally, data is publicly available and the relatively open access provided by Twitter, makes it a viable source for information diffusion and social network analysis.

Examples of analyses in the context of of information diffusion includes: how information does propagate on real time, how fast and for how long? Also, what are the users who play an important role in the information diffusion process? These results are interesting for scientists who study the social, temporal and structural aspects of information diffusion and end-users as well, who are engaged with visualization applications. Tracing the flow of information in real time offers significant insights on trending topics and emergency events.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Articles from this volume were invited to present their results at The 39th International Conference on Very Large Data Bases, August 26th - 30th 2013, Riva del Garda, Trento, Italy.

Proceedings of the VLDB Endowment, Vol. 6, No. 12
Copyright 2013 VLDB Endowment 2150-8097/13/10... $ 10.00.

Use case

We will highlight these analyses on an online journalism use case. Since social media has started to exert a big influence on the way that people are being informed, journalists cannot underestimate this powerful source. In order to take advantage of the rapid update rates and the huge amount of information and cover events while they are happening, journalists need fast and accurate analysis *on the fly*.

Journalists are involved in the diffusion process in three ways:

- It is relevant for journalists how information reaches them via the their followers lists, how long it takes and which are the intermediate steps. This results in an iterative process which validates and updates their sources.
- On the other side, after emitting some information they are interested to know what is the audience reached, how fast and what is the impact of it. Again, with an iterative process they can control and increase their impact.
- Lastly, they have to observe relevant trends and topics and they need the appropriate analysis and tools to increase efficiency.

In Section 2 we describe the problem and challenges, then in Section 3 we discuss the related work and in Section 4 we outline our ideas for tackling the problem. At the end, Section 5 concludes and gives future directions.

2. PROBLEM DESCRIPTION AND CHALLENGES

We plan to build an infrastructure for real-time analysis of social media with a focus on information diffusion. Our analysis part consists of fundamental steps that define information cascades out of the stream of messages and evaluations on top of them. The challenges we are going to face are multiple: (1) building an appropriate infrastructure that can handle both temporal and structural information in real-time, (2) implementing incremental and distributed algorithms for information diffusion and (3) implementing analysis with partially missing data.

Firstly, there is lack of systems and methods for *deep* and *fast* analysis of social media in real-time. Deep analysis includes complicated computational techniques such as modeling and predictive tasks which are implemented offline up to now. On the other hand, fast analysis is restricted mostly to lightweight computations, e.g. counting the number of top trending keywords over a period of time. Infrastructures that target Big Data provide only partial support for the requirements of real-time social media analysis.

The next fundamental step is to process the stream of messages that arrive in fast rates into a more structured and meaningful representation. Since our main topic is information diffusion, analysis is focused on reconstructing and defining information cascades, as described in Section 1. On top of this representation we can then perform more elaborate analyses.

Cascade reconstruction consisits of building subgraphs of the social network in order to model diffusion/influence paths of a stream of messages. As our initial analyses show (see Section 4.6) show, real-life cascade may reach to 100Ks of messages, while several thousands of cascades are active even in small subsets of Twitter. As a result, we target non-trivial, massive reconstruction of cascades on real-time. Up to now, such analysis has been performed in an ad-hoc manner, using offline, after-the-fact algorithms on small datasets [15].

The next fundamental issue we plan to address is to estimate how long each cascade remains active to determine is end time or perform predictions on its additional lifetime. As we observed on real-life data (and in line with other studies), cascade lifetime differs dramatically from some minutes up to weeks or months. Providing real-time results on complex cascade metrics with blocking computations require such an accurate estimation. Additionally, lifetime information help us to manage the large amount of cascade state.

Another significant challenge in tracking information diffusion in real data is incomplete datasets. This stems either from limitations in data availability of various providers and/or from increasingly privacy concerns of users who tend to keep their personal data constrained. As a result, when we study information propagation in social media, we end up with disconnected information cascades. The challenge lies at inferring information diffusion paths out of incomplete data. Moreover, since information cascades might be growing with rapid rates finding a trade off between missing data and quality of results is another challenge we plan to address.

After defining and reconstructing information cascades we plan to compute metrics and algorithms on them in a distributed and incremental way. Such analyses includes simple metrics, such as size and path lengths and complicated methods like identifying influential users. A more detailed description can be found in Section 4.4. These metrics offer interesting insights on information diffusion patterns and important users in this procedure. Up to now, such analysis is not implemented on real-time [27, 24, 7].

3. STATE OF THE ART

This thesis builds on work done on social media analysis, information diffusion and real-time analysis. We will briefly discuss related work in all of these areas.

3.1 Social media analysis

Regarding the social media analysis field, we are interested in the structural analysis of the underlying social network and stream analysis of the data flow. Social network analysis constitutes a long established field in many areas. Several scientists have investigated the structure of various social media platforms and implemented detailed analysis [32, 25, 44]. Normally, they study typical characteristics of social networks such as (among others) the existence of power law distributions and small world properties. In particular, regarding the Twitter social network, structural analyses show that it resembles more a news media than a classical social network, as reciprocity of connections is not high [25]. However, there is an underlying strongly connected social network beyond the declared follower and followee relationships over which most interactions take place [23]. Most of these analyses are implemented offline with an emphasis on the statistical properties of social networks while the temporal aspect is notably absent.

A more recent, complementary approach is to consider the activities on social media as a stream. We can discern two categories of stream-oriented analysis on social media: 1) streaming evaluation of static metrics and 2) dynamic, stream-oriented metrics. Regarding the first approach, given the large size of the data, it is clearly beneficial to perform the evaluation while the data is arriving over time. In this way, we minimizing state and response time: Schank and Wagner [38] designed an algorithm that can approximate clustering coefficients and transitivity in a streaming way in order to overcome the problem of massive graphs. Also, there is research that can approximate triangle counting in edge streams [43, 14]. The second approach embraces the streaming aspect also on a conceptual level, treating the social media as dynamic and time-oriented. Trend detection and event identification are two research areas that emerge as users become more and more involved in social media. ”Twitter monitor” [31] is a system that implements trend detection over the Twitter stream with the help of ”bursty” words. Similar but more targeted is the platform in [37] that identifies earthquakes on the fly out of the tweets stream. ”enBlogue” [4] is a real-time event identification platform that tracks shifts in correlations concerning tags and tag-pairs that identify emergency events. Interestingly, most of the research focuses purely on the temporal dimension, ignoring any dynamic structural aspects.

3.2 Information diffusion

The information diffusion field investigates how information (news, rumors etc) propagates among people. Most of the fundamental research on the flow of information and influence through networks has been implemented in the context of epidemiology and the spread of diseases [40]. We can discern two categories of related analysis which investigate 1) mechanisms of information diffusion and 2) structure and dynamics of information cascades as a more targeted approach. The first one includes how users share information and how different events impact on different temporal behaviour [45]. Moreover, identifying influentials or superspreaders is crucial for understanding the information propagation procedure [12]. Highly connected with the task of identifying key users is the problem of influence maximization which is how many and which users should be targeted in order to have maximum spread [13]. This problem is quite popular in advertising and viral marketing [26]. Another approach, ignoring completely the structure of the network is inferring paths of influence by taking into consideration only infection time [19]. Quite relevant to our project is the work by Ogan et al. [15] who study user interactions on Twitter and they are building algorithms to reconstruct the conversational graphs.

After observing information flow and especially information cascades, scientists implement various analyses: Lescovek [27] built a cascade generator, Zhou and Hui studied statistical, structural and content aspects of cascades [24, 47]. Cascade information is also used for predicting aspects of information diffusion e.g., which user will retweet which URL [46, 18]. Accurate cascade analysis is implemented when the full dataset is available. However, this is not always the case as discussed in the Section 2. Sadikov [36] built an algorithm that can infer cascades' properties out of sampled data. This contribution is quite valuable for social media scientists since full datasets are randomly available.

Until now, the work described including complicated analyses and offline implementations is aimed mainly towards scientists. There are also application specific lightweight approaches that target mainly the end user and work in realtime. ”Whisper” [10] follows a visualisation approach depicting how far information diffuses in real-time. Quite similar applications are ”Google ripples” [20] on Google+ and ”A world of tweets” including a heat map of tweets' activity [2].

3.3 Infrastructures and methods for big data management and real-time analysis

From a technical point of view, since we target complex computations on large amount of data on real-time, we need to explore the right types of infrastructures. On one side, we need a real-time system with the desired expressiveness and scalability. On the other side, we need graph databases in order to store and make computations on top of huge graphs. While we plan to implement our analysis in social media online, we still need to store instances of the social underling graph and be able to access it with good response rates.

The Map-Reduce framework has become synonymous with ”big data”. It provides simple abstractions as well as mature infrastructure to utilize very large computer clusters for computation and data management. However, it comes with a number of drawbacks. The restricted expressiveness and state modeling limit the suitablility of Map-Reduce for workloads with complex expressions and/or complex state. Moreover, the batch-oriented execution model and the focus on replication and recovery for high availability restricts efficiency and prohibits the use in scenarios with a strong temporal aspect.

”NoSQL” scalable datastores provide a good trade off in data and query expressiveness, consistency, architecture and scalability [11]. Two classes are relevant for our project: *Distributed key value stores* and *Graph Databases*. Distributed key value stores utilize a restricted set of operations on a minimal data model to achieve massive scalability, high performance and availability. They can offer a complementary storage when performing real-time evaluation since they are described by high performance but strict restrictiveness. An examples is HBase [5]. Graph Databases provide the means to store massive graphs in a distributed manner, and perform queries as well as complex computations on them. Similarly to key value stores, they still work with a store and process approach and do not provide much support for temporal analyses. Examples are Google Pregel [29] and GraphLab/GraphChi [28].

Data stream systems have been designed to deal with data that is continuously arriving and may be of infinite length. Data is treated as transient, and not stored. Queries on data streams, are considered as long-running, and their implementation will use techniques to produce results incrementally or on well-defined subsets such as sliding windows [8]. Core research areas cover model aspects such as time, ordering [17], operators [9, 6] and load management such as Shedding and Sampling/Synopses [16, 41]. Especially, distributed stream processing frameworks are extended to achieve better availability and scalability. Recently, a novel approach, inspired by Map-Reduce is gaining acceptance: Instead of providing a relatively closed, higher-level system like Borealis [3], frameworks like Storm [1] cover fundamental aspects such as the processing model, distribution, reliability and scheduling. The specific operator semantics and operations are defined by users or by separate higher-level projects.

To our knowledge, there is almost no research on approaches that are both ”deep” and ”fast”, since such an approach requires an understanding of scalable system design on top of the foundations of social media. We plan to use data stream systems that offer greater expressiveness and scalable algorithms in order to scale the huge datasets with high update rates.

4. RESEARCH IDEAS

In the following subsections we describe our ideas for building our infrastructure, developing relevant algorithms and tackling with the aforementioned problems. First, we refer to the fundamental analysis that extracts from the stream of messages information cascades and reconstructs them. Then, we describe metrics and algorithms on top of the information cascades that have been defined and reconstructed in the previous steps.

4.1 Reconstruction of complete cascades

In order to track information diffusion on real-time, we need to extract and represent information cascades out of the message stream. A cascade is formed when users forward the root's initial message, which can be stated explicitly by giving credit to it, for example with the retweet function. However, users do not always give credit to the initial contributor, either from choice or because it is not supported by the platform. In order to explore the second case, we need to track the propagation of messages e.g. URLs, taking into account the connections between users (social graph), timestamps, possible influence scores etc.

Cascades express the paths of information diffusion as directed graphs with a root node, since information can reach a participant via multiple ways. In the Twitter case, when a message is retweeted there is a reference only to the initiator of the cascade and not to the full path over which the message propagated. As a result, we are not just dealing with a graph construction problem, but also need to determine the most likely previous retweet steps. In order to do so, we plan to develop a model for the likely edge assignment including social graph aspects like distance (static) or influence strength as well as the timing of message flow (dynamic).

In order to provide low response times and scalability we aim for an incremental and distributed algorithm. The implementation of this algorithm will utilize the social graph snapshots for distributed storage and lookup. We will evaluate the algorithms with regard to result quality (including comparison models), reconstruction speed, resource utilization and resource consumption.

4.2 Detection and prediction of cascade boundaries

While trying to identify prerequisites for our analysis, for most metrics on information cascades identification of their boundaries is necessary. For example, specifying a cascade's duration or size and depth are metrics that require lifetime information. Cascades have no observable end condition since users can pick up a piece of information and forward it at any point in time. Cascades differ dramatically in terms of size, duration and activity, thereby we cannot apply the same ending condition for all e.g. after some hours of inactivity. Moreover, setting long durations contradicts the real-time concept of producing instant results. In addition, considering cascades "active" for an extended period creates an (unnecessary) state management problem, in particular when the full cascade contents are needed for analyses. Existing means for data stream management are not well suited to model cascade lifetime since they can only provide very coarse approximations of the boundaries. Cascade lifetime prediction methods yield usable results [18] but employ complex algorithms with significant amount of computation and state per cascade and are thus not scalable to the problem dimensions we are targeting.

Being aware of the aforementioned problems we will observe and identify indicators for cascade lifetime and develop scalable cascade boundary detection models out of them. The implementation of these models will be be evaluated for result quality, response time and resource consumption. In a secondary step, we plan to extend these detection models into lifetime prediction models by utilizing the indicators (or determining new ones) to forecast how much longer a cascade may last.

4.3 Reconstruction and metric computation on cascades with missing data

In the previous cases we take for granted that we are able to retrieve the full information cascade dataset. However, as already discussed this is not the case in reality due to privacy settings of users or due to various API's limitations. Twitter for example limits filter subscriptions to 1 per cent of its total stream, and requests to for follower lists to 60 per hour. Sadikov et al. [36] address the problem of estimating cascades properties having only a sample of the full data. Given that incompleteness occurs particularly often in trending, highly interesting cascades, we aim to design an incremental, distributed algorithm for reconstruction of incomplete cascades based on the ideas of the related work.

Given the huge amounts of data and the restrictiveness of state-of-the-art platforms to perform computations in real-time, we plan to determine evaluation models and thresholds for missing data so that still the results are acceptable. In addition, we will explore the trade off between completeness and efficient data management which is specific for each application.

4.4 Metrics and algorithms for cascades

After our fundamental analysis, reconstruction and determination of cascade boundaries are well defined. Our first goal is to analyse information cascades in terms of simple (counting) metrics, shapes, temporal aspects and more complex analysis like identifying influentials. We can discern the following categories for our proposed metrics covering many aspects of information cascades:

1. Basic measures like size in terms of users involved [7] [25, 47], diameter [27] and depth in terms of path lengths. The size of the cascade represents users who participate in the information propagation process by forwarding a specific message (tweet) of a root user
2. Metrics with regard to shape patterns like types [27], frequencies [27, 25, 47], correlations of shapes to events [24], degree distributions [27, 24]. Examples of shapes are stars, trees, long chains etc.
3. Metrics for temporal aspects like the time lag between messages (retweets) [25], cascade growth rate, that is how many messages of the same cascade (retweets) are being created in some time unit. A more complicated temporal metric is defining lifetime of cascades for which we plan to employ machine learning techniques in order to predict how long a cascade will last.
4. Collection of metrics that target specific users, in our case identifying influentials, in terms of structural factors like attribution of influence edges [7] and indegree [12] and temporal factors like past influence [7]. We plan to identify all the characteristics that render some users influentials which means that they can generate big cascades.

It is worthwhile to mention that basic measures mostly counting ones can approximate and summarize (sub)graphs. These metrics provide a wide range of analyses of information diffusion and offer insights in terms of how information diffuses structural and temporal wise. Especially when performed in a realtime manner, the aforementioned metrics can give preliminary information about the impact of events and potential influence of users and even perform predictions.

The current state-of-the-art lacks specific algorithms to compute these metrics in a real-time manner. We plan to implement algorithms or use existing offline ones that can approximate the aforementioned metrics. Since we target analysis in real-time we need to find ways to make them incremental and distributed.

Starting from relatively simple metrics on information cascades (e.g. size) and proceeding with more complicated ones (e.g. shape of cascades) we need to compute them in real-time i.e. an incremental and distributed way. For the more complex algorithms and models, only centralized, offline or batch-oriented algorithms are known [15]. We will analyse them, and utilize them as a baseline for our own algorithms. These algorithms need to be incremental and distributed.

Having designed and implemented algorithms to compute relevant metrics, we strive to improve their resource consumption, scalability and response times by interleaving reconstruction and

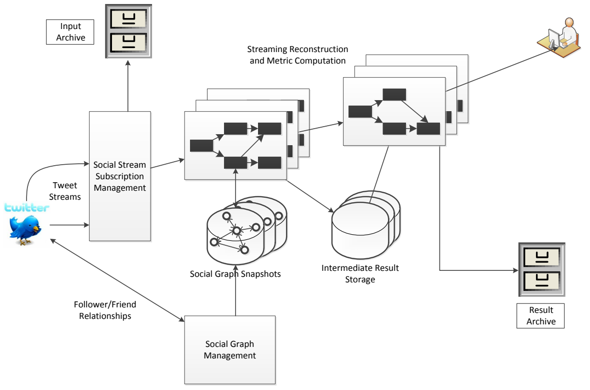
Input Archive
Streaming Reconstruction and Metric Computation
Social Stream Subscription Management
Tweet Streams
twitter
Social Graph Snapshots
Intermediate Result Storage
Follower/Friend Relationships
Social Graph Management
Result Archive

Figure 2: Infrastructure

metric computation. This interleaving not only provides performance benefits, but in the style of online query processing [35] allows us to produce results on partially observed cascades. We can therefore show preliminary results early, and improve them while the cascade is developing. In order to achieve this interleaving in an efficient manner, we need to devise intermediate representations such as synopses [39, 34] or graph summarizations [42, 33] to avoid maintaining full cascade graph state over long time.

4.5 Infrastructure components

Figure 2 shows the architecture diagram for our infrastructure, which we intend to instantiate as a prototype within this project. Generally speaking, the architecture aligns itself with typical architectures for large-scale real-time data processing: it combines components for the actual stream processing, pre-computation of relevant data as well as storage and access for materialized/precomputed and streaming data [21, 30]. Its main processing path is a scalable, distributed data stream processing platform, which we are currently evaluating. We use Storm [1] since it provides the necessary low-level primitives for distributed stream processing. Since we handle streaming data, we need state management components for the social graphs and the intermediate computation results. Our intermediate and pre-computed results are outcomes of crawling or stream processing, so we do not include a separate offline computation path using batch-oriented tools like Hadoop, as proposed in [30]. Besides the main computation path and storage, we need some supporting components to retrieve streams and social graph data, archive inputs and outputs.

4.6 Datasets and Preliminary results

Since we target information diffusion on Twitter, we have been collecting tweets from events having a global impact and possibly generate big cascades of retweets by subscribing to relevant hashtags. We have collected tweets from Olympics 2012 in London, US elections 2012 and Super bowl 2013. For example, the biggest cascade of retweets ever created was initiated from President Obama after winning the US elections which recorded over 800.000 retweets.

We have been exploring Storm [1] as a system for stream analysis and build operators on top of it in order to process the stream of tweets. For example, we have implemented selection, projection, count, sum, average, topk, window operators (centralized, distributed, grouping), I/O. At Figure 3 we can observe tweets and retweets during US elections. The spikes correspond to the debates and the last burst of tweets to the election day. The borderline of 3.000 approximately tweets is attributed to Twitter's limitation API of 1 per cent of the public stream. Figure 4 depicts the top-50 cascades in terms of size in 15 minutes time during the Olympics. We can observe a skewed distribution of cascades' size, however most of them are quite big. On the other side, cascades' lifetime differs massively. Observations from our dataset show that they can span from days up to years, in extreme cases.

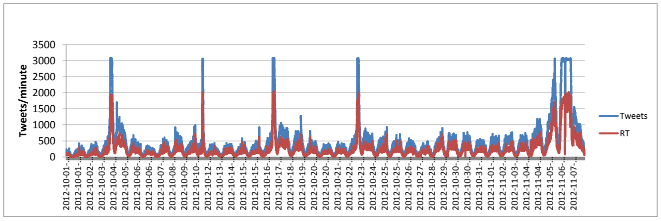

Figure 3: Tweets and Retweets for 38 days of US elections data

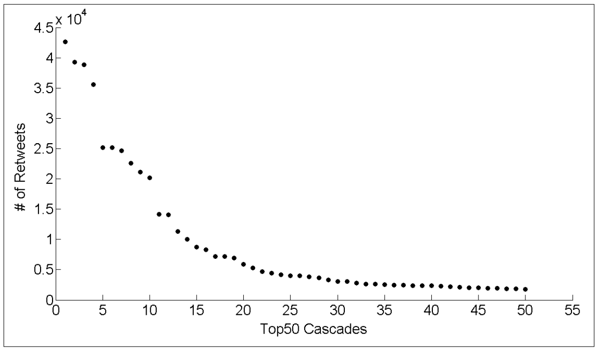

Figure 4: Top50 largest cascades in 15 mins time of Olympics data

5. CONCLUSION

The increasing usage of social media provides researchers with many opportunities to study human interactions such as social relationships, development of communities or information diffusion. The large amount of data provides the means to study these phenomena on representative user groups; the rapidness of interaction gives way to real-time indicators or even predictions.

A particularly interesting field for real-time analysis is information diffusion, analysing and predicting how information spreads. Research to understand and utilize information diffusion follows typically two major directions with little intersection: On the one hand, there are sophisticated models that describe and approximate information flow in social media. Given their complexity, they are applied on small, often sampled data in an offline fashion, targeting scientists. On the other hand, there are lightweight analyses with visualizations that are computed in real time and target end users.

As far as we are aware, current literature lacks from systems and algorithms that perform complex and incremental analysis in real time. This thesis aims is to fill this gap by providing an infrastructure for real time analysis in huge datasets containing structural information (graphs). Even if the project described is in a very early stage, we are confident that the contributions will be useful for the scientific community and the end users as well. Moreover, the convergence of the fields of social media analysis and real-time systems offers novel and challenging ideas for research.

6. REFERENCES

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

[2] A world of tweets, 2010. http://aworldoftweets.frogdesign.com/.

[3] D. J. Abadi et al. The design of the borealis stream processing engine. In *CIDR*, pages 277–289, 2005.

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

[5] Apache Foundation. Hbase. http://hbase.apache.org/.

[6] B. Babcock et al. Models and issues in data stream systems. In *PODS*, pages 1–16, 2002.

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

[8] I. Botan et al. Extending xquery with window functions. In *VLDB*, pages 75–86, 2007.

[9] I. Botan et al. Secret: A model for analysis of the execution semantics of stream processing systems. In *PVLDB*, pages 232–243, 2010.

[10] N. Cao et al. Whisper: Tracing the spatiotemporal process of information diffusion in real time. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2649–2658, 2012.

[11] R. Cattell. Scalable sql and nosql data stores. In *SIGMOD Record*, pages 12–27, 2010.

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

[13] Y.-C. Chen et al. Efficient algorithms for influence maximization in social networks. In *Knowl. Inf. Syst.*, pages 577–601, 2012.

[14] S. Chu and J. Cheng. Triangle listing in massive networks and its applications. In *KDD*, pages 672–680, 2011.

[15] P. Cogan et al. Reconstruction and analysis of twitter conversation graphs. In *Interdisciplinary Social Networks Research*, HotSocial '12, pages 25–31, New York, NY, USA, 2012. ACM.

[16] G. Cormode et al. Continuous sampling from distributed streams. In *J. ACM*, page 10, 2012.

[17] A. J. Demers et al. Cayuga: A general purpose event monitoring system. In *CIDR*, pages 412–422, 2007.

[18] W. Galuba et al. Outtweeting the twitterers - predicting information cascades in microblogs. In *Online social networks*, WOSN'10, pages 3–3, 2010.

[19] M. Gomez Rodriguez et al. Inferring networks of diffusion and influence. In *SIGKDD*, pages 1019–1028, 2010.

[20] Google. About google+ ripples, 2013. http://support.google.com/plus/answer/1713320?hl=en.

[21] M. Grineva et al. Analytics for the realtime web. In *PVLDB*, pages 1391–1394, 2011.

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

[23] B. A. Huberman et al. Social networks that matter: Twitter under the microscope. In *CoRR*, 2008.

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

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

[26] J. Leskovec et al. The dynamics of viral marketing. In *TWEB*, 2007.

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

[28] Y. Low et al. Distributed graphlab: a framework for machine learning and data mining in the cloud. *Proc. VLDB Endow.*, 5(8):716–727, Apr. 2012.

[29] G. Malewicz et al. Pregel: a system for large-scale graph processing. In *SIGMOD Conference*, pages 135–146, 2010.

[30] N. Marz. Big data lambda architecture, 2008. http://www.databasetube.com/database/big-data-lambda-architecture/.

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

[32] A. Mislove et al. Measurement and analysis of online social networks. In *Internet Measurement Comference*, pages 29–42, 2007.

[33] S. Navlakha et al. Graph summarization with bounded error. In *SIGMOD Conference*, pages 419–432, 2008.

[34] N. Polyzotis and M. N. Garofalakis. Statistical synopses for graph-structured xml databases. In *SIGMOD Conference*, pages 358–369, 2002.

[35] V. Raman and J. M. Hellerstein. Partial results for online query processing. In *SIGMOD Conference*, pages 275–286, 2002.

[36] E. Sadikov et al. Correcting for missing data in information cascades. In *WSDM*, pages 55–64, 2011.

[37] T. Sakaki et al. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, pages 851–860, 2010.

[38] T. Schank and D. Wagner. Approximating clustering coefficient and transitivity. In *J. Graph Algorithms Appl.*, pages 265–275, 2005.

[39] X. Shi et al. The very small world of the well-connected. In *Hypertext*, pages 61–70, 2008.

[40] E. Stattner and N. Vidot. Social network analysis in epidemiology: Current trends and perspectives. In *RCIS*, pages 1–11, 2011.

[41] N. Tatbul et al. Load shedding in a data stream manager. In *VLDB*, pages 309–320, 2003.

[42] Y. Tian et al. Efficient aggregation for graph summarization. In *SIGMOD Conference*, pages 567–580, 2008.

[43] C. E. Tsourakakis. Fast counting of triangles in large real networks without counting: Algorithms and laws. In *ICDM*, pages 608–617, 2008.

[44] J. Ugander et al. The anatomy of the facebook social graph. In *CoRR*, 2011.

[45] S. Wu et al. Who says what to whom on twitter. In *WWW*, pages 705–714, 2011.

[46] Z. Yang et al. Understanding retweeting behaviors in social networks. In *CIKM*, pages 1633–1636, 2010.

[47] Z. Zhou et al. Information resonance on twitter: watching iran. In *Social Media Analytics*, SOMA '10, pages 123–131, 2010.