Mining the Boundaries of Social Networks: Crawling Facebook and Twitter for BlogIntelligence

Philipp Berger[1], **Patrick Hennig**[1], **Thomas Klingbeil**[2], **Matthias Kohnen**[2], **Steffen Pade**[2], **and Christoph Meinel**[3]
Hasso-Plattner-Institute, University of Potsdam, Germany
[1]{philipp.berger, patrick.hennig}@hpi.uni-potsdam.de
[2]{thomas.klingbeil, matthias.kohnen, steffen.pade}@student.hpi.uni-potsdam.de
[3]office-meinel@hpi.uni-potsdam.de

Abstract— *Today's number of weblogs is higher than ever before and still growing. These blogs are interconnected by numerous links and other diverse connections, generating a series of notable patterns. Weblogs are not isolated and highly connected with other social networks like Facebook and Twitter. Thus, we analyze the references and investigate methods to gather data from the social platforms that are interconnected with weblogs. By analyzing the communication flow between weblogs, Facebook and Twitter, we observe that Facebook is mostly used for referencing real people instead of posts. In contrast, tweets are primarily used for information propagation and citation.*

Keywords:
Data Mining, Social Networks, Weblogs, Twitter, Facebook

1. Platforms in the Social Web Have to Be Connected

Weblogs, called *blogs*, are one of the most popular "social media tools" of the World Wide Web (WWW) [1]. They are specialized, but easy-to-use content management systems. Blogs focus on frequently updated content, social interactions, and interoperability with other Web-authoring systems.

The actual power of blogs evolves through their common superstructure, i.e. a blog integrates itself into a huge think tank of millions of interconnected weblogs, called blogosphere that creates an enormous and ever-changing archive of open source intelligence [2].

The structure of the whole social web has undergone a huge shift within the last years. Instead of using a single social platform users tend to use multiple platforms in parallel.

Today's social web consists of a collection of these platforms like Facebook[1], news portals, weblogs and diverse other intercommunication websites like *Twitter*[2], *Pinterest*[3], and *Foursquare*[4]. Research around social networks focuses

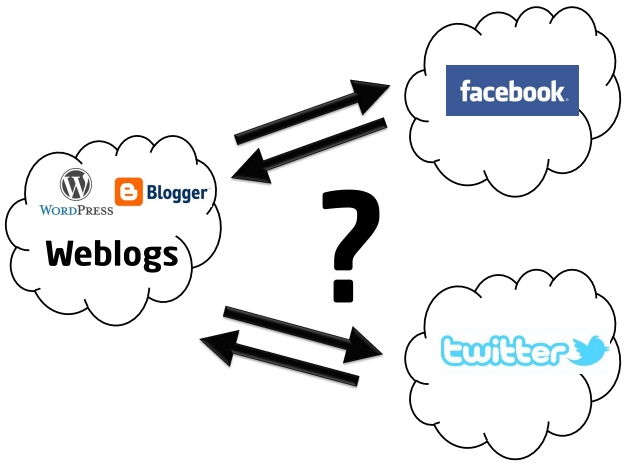

Fig. 1: How are weblogs and other social networks connected?

on one specific platform and investigates the communities, information flow, and social structure within this platform. One example is the BlogIntelligence[5,6] project. Within the scope of this project there have been several research efforts on structure, growth, and emergence of weblogs. Although blogs account for a major segment of the social web, different analyses show that weblogs extensively link to other social networks, especially Twitter and Facebook.

We observe that besides linking social profiles, bloggers use external platforms to announce posts, redirect discussions (instead of using comments), pickup controversial opinions or reference people. These observations in mind, we identify the need for a deeper analysis. Therefore, we need to collect data from these "external" sources, first, and secondarily put them into a semantic relation to the already gathered data from weblogs.

This leads to various new insights and at the very least offers a new perspective on how connections between the

[1] http://facebook.com
[2] http://twitter.com
[3] http://pinterest.com
[4] http://foursquare.com

[5] http://blog-intelligence.com
[6] http://hpi.uni-potsdam.de/meinel/knowledge_tech/blog_intelligence

different kinds of social web platforms are created and maintained (see Figure 1). Interesting research questions on this topic are for instance the differences in user activity between the platforms or how trends spread among them. The results of this research include analyses on how topics (especially high interest, popular trending topics) spread among the platforms or whether or not the platforms concentrate on different fields of topics. Are users of two or more of the platforms also talking about the same things on all of them? This and many more questions could be answered by these results. More details on what could also be of interest will be given in Section 5.

Within the scope of this paper we investigate methods and realize harvesting applications for Facebook and Twitter. Since BlogIntelligence is only focused on weblogs, we need to extract the connections to other social platforms from the existing data set to find links pointing to Facebook and Twitter.

The next section gives an overview of related work. Sections 3.1 and 3.2 focus on the crawling processes for Facebook and Twitter. The data retrieved by these processes is then analyzed in Section 4. This paper closes with recommendations for further research in Section 5 and a conclusion in Section 6.

2. Related Work

We distinguish two areas of related work. First, related approaches for harvesting the social networks in scope, e.g. Facebook and Twitter. Second, approaches towards mining of the interaction between social networks.

Under the name *TwitterEcho*, a research group has already developed an open source Twitter crawler [3]. Their work is using the REST API, as Twitter still allowed whitelisting during the time of their research, in order to increase the allowed number of requests per hour. As whitelisting is no longer possible, we had to focus on finding an algorithm, which intelligently uses the Streaming API to achieve our goals.

Another group from INRIA Sophia Antipolis has focused on acquiring a full overview of the user base of Twitter and drawing a graph of the way the users are connected [4]. For their work, they used a distributed platform, called PlanetLab. Their findings include information regarding user activity and the influence of Twitter policies and social conventions on the structure of that graph. In contrast to our work, their gathered data does explicitly not contain the content of the tweets.

The topic of the topological characteristics of the Twitter network has also been picked up by H. Kwak et al. from the Department of Computer Science, KAIST, Korea. They compared the way social networks work to characteristics of traditional human social networks. This research group also introduced a PageRank ranking algorithm for Twitter users. As a result of their research they presented that over 85% of Twitter posts are news-related content [5]. For us this means, that linking the information from Twitter to the information already gathered about the Blogosphere is an important step.

Apart from Twitter, also Facebook crawling has been conducted by other researchers at the University of Messina, Italy [6]. Again, they looked into the details of the connections and interactions between the participants of Facebook. They have used Breadth-first-search sampling, which means seed nodes have been employed at the beginning of the crawling phase. Instead of using the faster Graph API provided by Facebook, they used the deprecated Ajax interface.

Another important topic which needs to be considered is, how spam users can be detected and filtered from the data to be analyzed. Research in this field has been conducted by F. Benevenutu et al. from the Computer Science Department of the Universidade Federal de Minas Gerais Belo Horizonte, Brazil [7]. They presented an algorithm, which allows a precision of 70% while classifying spammers, which is based on detecting specific characteristics using machine learning techniques.

To the best of our knowledge the research concerning cross-platform social media mining experiences only little investigation in the community. Quandt et al. [8] relate social networks and traditional channels like newspapers and discuss the opinion towards the quality and usefulness of weblogs for journalism. Likewise, Hermida et al., [9] investigate the interaction between television and Twitter. From an architectural point of view, Pallis et al. [10] dive into the similarities of social networks with the goal to develop cross-platform services.

In contrast to related work, we explore the relations across online social networks and try to identify unknown connections, diffusion mechanisms, et cetera.

3. Social Network Harvesting

To mine different social networks and their boundaries we need to harvest the publicly available information and make them available offline. This enables us to run offline cross-network analyses. As mentioned above, the BlogIntelligence project already stores blog data that a tailor-made crawler downloads. Thus, we focus on finding referenced social networks in this data set and on developing adapted crawlers for Facebook and Twitter.

3.1 Facebook Crawling

Within the last years Facebook has grown to be the world's largest social network according to their active user groups. It has been of no surprise that the BlogIntelligence data contains a high number of references to Facebook pages, posts and user pages. Due to the structure of the filtered links and the structure of the social network itself, it needs to be considered that there are different instances of Facebook entities. Whereas the individual user pages are well-known, there are also pages of businesses, celebrities, groups and

diverse other information sources. Due to the API, which will be explained in more detail in the next part of this section, the data preprocessing needs to identify these types of pages and decide how to crawl the source.

Further, the somewhat unclear terms of use of this API have led to uncertainty on what exactly an automated web crawler is allowed to do with Facebook. Since this uncertainties could not be cleared so far we have decided to continue gathering data in a smaller and cautious way by only running the application occasionally with lighter sets of data.

3.1.1 API and Restrictions

Although there is an old legacy REST API and the FQL (Facebook Query Language) API the only reasonable interface to use is the Facebook Graph API which allows for the execution of RESTful requests with JSON formatted data against their website. The structure of such a query is held quite simple and makes exact requests even for a single post possible.

With that in mind the filtered links pointing to Facebook had to be investigated on what exactly they are pointing to - a page, a user, etc. - and the according request had to be made. Unfortunately the terms and conditions[7] applying to the usage of the Facebook Graph API do only consider using this API for building a so-called Facebook app. This term describes an application that is either a stand-alone product connecting and authenticating a Face book user or a web application that can be accessed via the Facebook website. The automated data collection has only the requirement that the collector has to make the collected data searchable, collect data for purposes of search respectively. Since we will be able to visually represent the gathered data in our webportal we thus comply to this requirement by our integrated search functionality.

3.1.2 The Data Collection Process

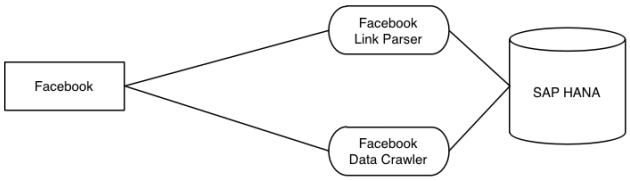

Fig. 2: Conceptual view on the Facebook crawling process

The data collection process consists of four steps:

1) filter and normalize links from BlogIntelligence data
2) select most active users
3) connect to Facebook
4) download and store information

[7] http://www.facebook.com/apps/site_scraping_tos_terms.php

a) Step 1: This step is the most challenging caused by the limited amount of requests and the relatively noisy data set. Hereby, we need to filter, prepare and understand the links within the BlogIntelligence data set.

As stated before, Facebook as a social network introduces different types of entities like users, groups, posts, and events. Since the focus of BlogIntelligence is mainly on tracking discussions of blog authors on the internet, the focus of this project was placed in the same manner. For that reason Facebook events and groups are excluded from the collection process. These URLs were filtered out as well as all the corporate Facebook pages like help pages, information pages and obviously the home page itself.

By analyzing the given set of links we empirically detect link patterns to distinguish between user pages and unusable links. Further, we normalize those links by removing all but the user identification alias. This excludes unwanted subpages and allows us to easily match links of different posts to the same user. The user alias is used to request all the posts of one user for a given time frame. Since there are no restrictions to the number of requests, this process can run in parallel for each user every day or even every couple of hours. Nevertheless, bandwidth and server time restrictions can dictate us to concentrate on a subset of all Facebook users.

b) Step 2: During our test period we ran this process two times within 30 days. Moreover, these two executions provided the chance to research user activity. This leads us to the prioritization of users depending on their activity on Facebook.

Therefore, we distinguish between "very active" and "less active" users by incorporating the posting frequency of users. We rank users according to their activity and only queue the top-k users for regularly updating.

To react to changes in user behavior only the last 30 days are taken into account. The evaluation of user activity is executed before every harvesting stage. With a growing dataset the ranking of users gets more and more accurate. Especially during the initial crawling the number of posts crawled for each user is quite low and there are also many users without any crawled posts. To deal with this issue the harvesting is restarted after quite short time frames to enable the collection of data for all users.

This distinction method works more satisfactory if a bigger number of posts is crawled and if for most of the users posts have been found.

c) Step 3: The connection to Facebook is mainly handled by an external library called *restfb*[8]. It encapsulates Facebook's Graph API into an easy-to-use Java interface. This library especially simplifies the authentication with Facebook.

[8] http://restfb.com/

d) Step 4: This step consists of downloading and storing the received data. Thus, we use restfb to request the wanted data. We translate the data into our own structures and store it using JDBC drivers into our relational database called *SAP HANA*[9].

3.2 Twitter Crawling

Regarding the number of active users, Twitter is Facebook's largest competitor. The BlogIntelligence data set reflects this by including also a high number of links to Twitter. Based on our observations, we like to crawl Twitter users because we assume that most of the Bloggers also maintain a Twitter account for publishing their posts and additional ideas. Nevertheless, the crawling of tweets of these users and of additional linked content is the logical next step.

The opportunities offered by Twitter's APIs are described in the next part of this section. Following this, we reflect on the implementation of the crawling process.

3.2.1 APIs and Restrictions

Twitter offers two different APIs with specialized capabilities. The *REST API* enables the access to all Twitter resources like tweets, user information, followship graphs and many more. The *Streaming API* provides the ability to retrieve a continuous stream of tweets for selected users.

During the year 2012 Twitter launched a new version of both APIs changing restrictions and methods. The old REST API version 1 allowed 350 requests to the REST API per hour and authenticated user. The new version 1.1 distinguishes between different REST API methods and restricts the number of requests depending on and per called method for each authenticated user. For some methods the restriction is set to 15 requests every 15-minute window. Nevertheless, the methods used by our Twitter crawler only have a restriction of 180 requests per 15 minutes. This allows 720 requests per hour with the API methods of version 1.1, which is more than twice the amount available with version 1.

Caused by this restriction we develop a method to combine both APIs to enable the best crawling performance by sticking to the restriction.

After identifying the Twitter user accounts within the BlogIntelligence data set, we use the REST API to gather all past tweets for this users. So, the REST API helps us to fill our tweet archive and collect the initial seed of tweets. Furthermore, this API is of high interest for time-discrete crawling of less active users. Thereby, we avoid to idle while waiting for new tweets of these users.

The coverage of highly active users tweeting many times a day or even per hour is very expensive in terms of requests to the REST API. This is exactly where the Streaming API is of crucial value. We use the full capacity of this API to observe our most active users. This enables us to continuously collect each new post of these users. The distinction into "less active" and "highly active" let us use each API to its limits and the crawler can gather tweets of identified users in the least possible time.

3.2.2 The Data Collection Process

The collection process of Twitter is similar to the Facebook crawling and consists of the same 4 steps (see Section 3.1.2):

1) filter and normalize links from BlogIntelligence data
2) select most active users
3) connect to Twitter
4) download and store information

a) Step 1: We empirically identify patterns for the Twitter link recognition. Hereby, we have to distinguish between links containing the screen name and links containing the user IDs. The screen names are human readable and necessary for user interfaces like a webportal. The user IDs are required for accessing the above mentioned APIs to crawl tweets, which are not directly linked.

b) Step 2: Within the process of crawling tweets from Twitter users, the same approach for distinguishing active and less active users for Facebook crawling is applied. This process was described in Section 3.1.2.

In contrast to Facebook, Twitter offers the Streaming API that enables us to get more frequent updates for a user group with a limited size of 5 000. Thus, for the most active users we use the Streaming API to retrieve continuous up-to-date tweets. The REST API allows for crawling tweets of the less active users. We conclude from the less frequent posting activity in the past that these users will continue to post in an infrequent manner. Thus, the REST API is sufficient to retrieve all tweets, even with the described limitations.

c) Step 3: To access the Twitter API we use *twitter4j*[10].

d) Step 4: Besides differences in the data structure, the storing process is the same as for Facebook.

4. Data Analyses

In this section, we show our first analysis results that directly result from a real life crawled data set obtained by BlogIntelligence. First, we present the key indicator of the base data. Following, the insights into the data collected from Facebook and Twitter.

[9] http://www.sap.com/hana/

[10] http://twitter4j.org/

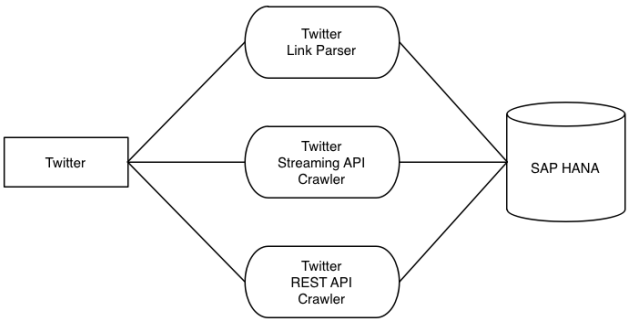

Fig. 3: Conceptual view on the Twitter crawling process

4.1 BlogIntelligence Data

The used data set of BlogIntelligence for this evaluation consists of 15 327 blogs with 818 865 posts. These posts consist of 200 000 000 links. This data set is the result of a 3-week-run from August 2012. Since we have integrated our crawling components into the BlogIntelligence Framework, we use this data as a basis.

4.2 Facebook and Twitter Data

We identify 554 962 links to Facebook users or posts of Facebook users. There are 31 825 distinct links. This implies that most users are linked multiple times. On average each user is referred to 17.4 times where 28 292 unique users can be extracted from the links.

The usage of Twitter is quite different because the 325 659 distinct links to Twitter users or tweets occur 1 425 244 times. Each link is used less often than Facebook link posts on weblogs (on average 4.4 times). Furthermore, only 13 589 unique Twitter users can be parsed from the links. A deeper analysis of the links shows that many times tweets are linked which are related to the topic of weblog posts. These numbers support the logical inference that links pointing into Facebook's social network are supposed to link to a Facebook user's profile and mainly inform about the existence of such a profile. Whereas, the linkage of tweets seems to link to a third-party source of information or to refer to a citation. This is also an indicator for the transient nature of tweets.

Posts and tweets crawled from Facebook and Twitter can also be analyzed to learn more about the structure of the data. Due to the afore-mentioned legal problems regarding the crawling of Facebook the main focus of this analysis is based on Twitter data.

From Facebook 96 317 posts were crawled during the short crawling periods. 64 353 of these posts contain a link to an external resource. Thus, over 60% of Facebook posts connect to other webpage that bring up new questions like "Do these links point to other weblogs?". Nevertheless, for more detailed insights this data set is too limited. Thus, we need to continue crawling to run deeper analyses.

Twitter was crawled for two weeks in February 2013. During this time 3 760 577 tweets were retrieved.

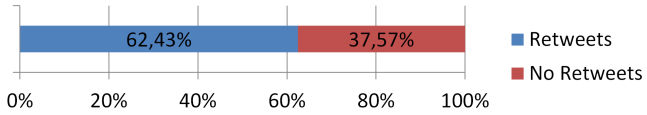

Fig. 4: Ratio of tweets being retweets to own tweets

One main feature of Twitter is the ability to retweet someone's tweet. This enables users to spread information and likewise show their appreciation for a tweet. As shown in Figure 4, 62.43% of all tweets are retweets of other tweets. This means that 2.3 million tweets are created just by the retweeting of an original tweet. The set of retweets references 216 911 tweets. The maximum retweet count of a single tweet in the data set is 30 888. This indicates the popularity of Twitter as a publishing channel that can rapidly spread information through its whole user base.

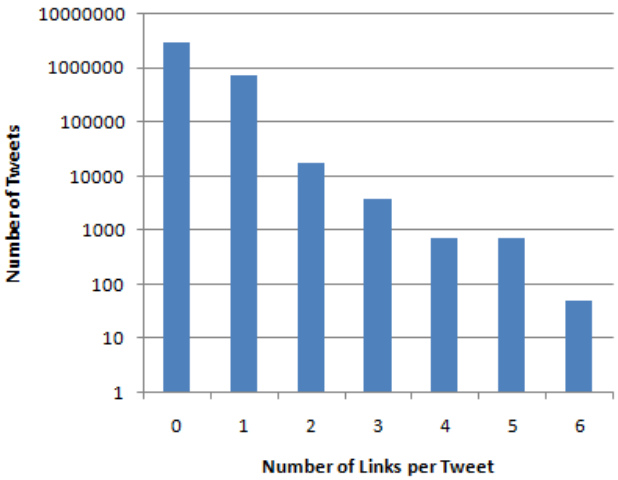

Fig. 5: Number of links occurring in one tweet, please mind the logarithmic scale

Many tweets are used to refer to other external websites by links. The analysis of the tweets, depicted in Figure 5, points out that most tweets, about three million, contain no links. About 740 000 tweets contain one link. The number of tweets with more than one link is quite small. This is also characteristic for the short messages of Twitter, but also indicates that the information flow does not stop in the referenced social network. It may also be doubtful whether tweets containing up to six different links have any relevant content.

Hashtags allow the Twitter users to give a short summary of the content of a tweet. This feature is widely used when posting about an event or discussing ongoing topics. This allows many different analyses to be performed on the data

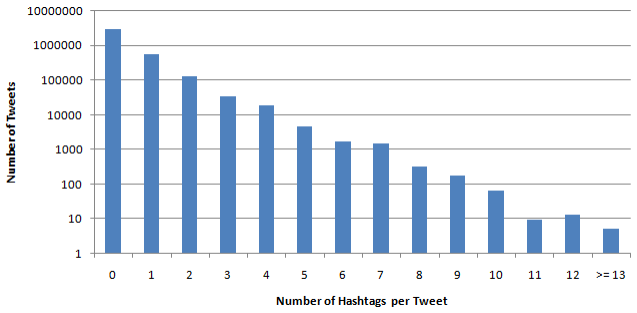

Fig. 6: Number of hashtags used in one tweet, please mind the logarithmic scale

like trend detection or clustering. The evaluation of the Twitter data regarding hashtags is quite similar to the analysis of links in the tweets, about three million tweets contain no hashtags, 550 000 tweets have one hashtag and two hashtags are assigned to 130 000 tweets. A higher number of hashtags is used less frequently and again it is questionable whether up to 18 hashtags contain any relevant information. The distribution of hashtags is shown in Figure 6.

5. Recommendations for Further Research

As described in Section 1 this project's goal was to implement the first step of a larger whole. Thus, only a limited result set is presented. The next step will be a more dedicated set of analyses on how weblogs and weblog networks interact with Facebook and Twitter by representing other types of social networks.

These analyses should aim to answer different questions of interest. In the scope of the present research activity within this project a next step will consider the relations between blogs, blog posts and the gathered tweets and Facebook posts respectively needs to be determined, which will then provide possibilities for investigating connections between the networks.

At this point there will be different fields to be regarded. A first interesting point will be how topics spread among the blogosphere and social networks when considering time and intensity. Upcoming interesting questions are:

- What time-gap lies between the first appearance of a topic and its encroaching to platforms of other types?
- Is there a platform (blogosphere / Facebook / Twitter) where new topics mostly appear the first time?
- Are main topics of one platform also main topics of all / one other platform?
- Which users are actively posting on all of the platforms?
- Which platforms are best synchronized concerning their main topics?
- Is a user who is active on two or more distinct platforms talking about the same things on all of them?

Answering these questions will give first indications on correlations between the blogosphere and Facebook and Twitter based on discussed topics and possible common or distinct user groups. Regarding the activities of users next steps will include some kind of an activity index calculated on possibly how often a user makes a post on one of the platforms or an activity index calculating how often new posts are made all over on one platform. These indexes in turn can give a metric to compare the platforms based on the activity of their members.

Furthermore, the gathered data from the social networks contains further links pointing to network internal and external resources. At this point the unanswered question is how much sense it would make to follow these references and include what they are pointing to into the data collection process. Whereas it was not considered for this project so far it might lead to new and more specific insights. When taking the second level links into account it will be necessary to distinguish them from the first level links originating in the blogosphere, especially their relevance to topic determination, trend detection or user activity. They have to be put into relation by weighing their importance against the large whole.

As a last point it should be mentioned that there are several other social networks out there. Since the follow-up step of this research project will also take them into consideration. The more data sources there are the more interesting and significant this project's results will be in the future.

6. Conclusion

We introduced the area of cross-platform social network analysis. Our work is based on the BlogIntelligence project and thus our starting social network is the blogosphere. By investigating the link structure of blogs we found numerous connections to other social networks especially Facebook and Twitter.

To investigate these connections we implement a harvesting application for both networks that makes the relations available for analyses. We conclude that even though gathering the data itself is easy, as comprehensible APIs are available from the providers, a lot of legal aspects need to be considered. Amongst others, this concerns the collection of personal data of users which even though publicly available, undermines certain rules. Additionally, the providers restrict the amount of data which can be retrieved within a specified amount of time. This makes it necessary to create intelligent algorithms which specify which data will be fetched at which point of time.

As a preliminary result of our research, we deduct that weblogs are strongly interconnected with the social networks Twitter and Facebook. These connections are bi-directional, as on the one hand blog posts are linked in Twitter and Facebook and on the other hand, weblog authors write about the content of tweets and Facebook pages. This advanced

level of relationship analysis can lead to the creation of a whole new *meta network*, interconnecting parts of the traditional blogosphere and social networks.

We analyzed the characteristic of the connected Facebook links and observed that those are mostly used for referencing people instead of posts. In contrast, the Twitter links mostly refer to tweets and we observe that these tweets are primarily used for information propagation.

References

[1] T. Cook and L. Hopkins, "Social media or, "how i learned to stop worrying and love communication"," September 2007. [Online]. Available: http://trevorcook.typepad.com/weblog/files/CookHopkins-SocialMediaWhitePaper-2007.pdf

[2] J. Schmidt, "Weblogs: eine kommunikationssoziologische studie," 2006.

[3] M. Boanjak and E. Oliveira, "TwitterEcho: a distributed focused crawler to support open research with twitter data," *Proceedings of the 21st ...*, pp. 1233–1239, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2188266

[4] M. Gabielkov and A. Legout, "The complete picture of the Twitter social graph," *Proceedings of the 2012 ACM conference on ...*, pp. 20–21, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2413260

[5] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter , a Social Network or a News Media?" *... of the 19th international conference on ...*, 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1772751

[6] S. Catanese, P. D. Meo, and E. Ferrara, "Crawling facebook for social network analysis purposes," *arXiv preprint arXiv: ...*, pp. 0–7, 2011. [Online]. Available: http://arxiv.org/abs/1105.6307

[7] G. Magno and T. Rodrigues, "Detecting Spammers on Twitter," *Science*, pp. 1 – 9, 2010. [Online]. Available: http://www.nber.org/chapters/c2665

[8] T. Quandt and J. B. Singer, "Convergence and cross-platform content production," *Handbook of journalism studies*, pp. 130–144, 2009.

[9] A. Hermida, "From tv to twitter: how ambient news became ambient journalism," *Media/Culture Journal*, vol. 13, no. 2, 2010.

[10] G. Pallis, D. Zeinalipour-Yazti, and M. D. Dikaiakos, "Online social networks: status and trends," in *New Directions in Web Data Management 1*. Springer, 2011, pp. 213–234.