Internal report 2012--02

Universiteit Leiden

Computer Science

Topical Influence on Twitter:
A Feature Construction Approach

Name: Menno Luiten
Student-no: 0345296

Date: 28/05/2012

1st supervisor: dr. W.A. Kosters
2nd supervisor: F.W. Takes MSc

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Topical Influence on Twitter: A Feature Construction Approach

Menno Luiten
LIACS, Universiteit Leiden

May 28, 2012

Abstract

Social media activity is increasing at an impressive rate. More than ever, companies and scientists realize there is a lot of value hidden inside the huge amounts of data, one of which is user influence. This research aims to provide an evidence-based metric on topical influence to identify conversation leaders and influential intermediaries based on the content of their messages. We do this by generating features based on the social graph and content of messages, and investigate the relation with several goal attributes. We then use this information to find an easily interpretable measure of topic-based influence.

Acknowledgements

I would first and foremost like to thank my supervisors, Walter Kosters and Frank Takes, for taking the time and effort to go through the many revisions of this research and steering my focus. Also, a big thanks to the helpful people in LIAAD, University of Porto, for showing me to be open-minded and allowing me to explore the many interesting things that surround this field of research. Finally, I would like to thank my fiancée, Manon, my family and friends for their inexhaustible support.

Contents

Chapter 1

Introduction

In this chapter we will describe the motivating theories for this research and give a general introduction to the research that has been done on influence theories in several fields. We will also explain the structure of this research and explain the scope of our approach.

The concept of how messages, concepts and information spread through human institutions has been subject to research for a very long time, ranging from the early philosophers to more advanced (or at least documented) studies in the last century. It has been examined in the fields of sociology, communication, marketing and political science. Its dynamics explain how societies function as a whole, but also how information flows on a lower, more direct level. It explains the function of the smaller agents in the larger ecosystem of society, and how every action has a reaction of different size. Most interestingly, it might also explain how to *control* the system. In marketing and advertising specifically, prior to the 1950s the production concept was commonly used; it was the belief that when goods are widely available and cheap, people will buy them. It was a concept in which everyone was equally influential, or even did not play a role in the decision making process.

In a study done by Katz and Lazarsfeld [23], it was shown that in many situations, information only reaches the majority of the people through opinion leaders, who in turn receive their information from media. This concept is called the *two-step flow of communication* model. These opinion leaders are portrayed to be the large cogs, influencing the decision making process of smaller cogs. It introduced the term "personal influence", meaning the opinion leader's ability to intervene between the media's message and the opinion of the majority. One possible effect of this research was the increased use of known personalities in messages of both political as commercial parties.

In marketing, the term *market mavens* was only recently introduced [15, 9], as a distinction from "connectors" and "salesmen": a group of people we trust with giving us new, valuable information in a certain area. In the advertising business, the recognition of the existence of these "trust centers", led to an increase of advertising based on well-known figures. Actors, singers and sports

players were often used as brand associations, hoping to influence the many. However, it was also clear that the message of these mavens is not spread directly from them to the millions who receive and/or adopt it. Rather, there is a large chain-reaction of intermediaries that spread the message through their respective word-of-mouth networks.

In recent years, more modern studies have shown that marketing mavens are not as influential as originally thought [35]. Instead, influential messages are adopted mostly based on the word-of-mouth network of moderately connected people as well as the *content* of the message [43]. This can also be seen in the quite recent marketing trend of *evangelism marketing* in which companies try to build such a strong relationship with their customers, that the customer becomes a voluntary advocate of the company's products. We can also explain the rise of review websites and recommender systems from these theories. All these strategies use the influence of the "reasonably influential" instead of the "marketing mavens".

So what is it about the moderately connected that makes them influential? In research by Nielsen [44] 90% of the respondents trust a recommendation from "people they know", an increase of 12 percentage points from 5 years earlier. 70% of respondents also say they trust "consumer opinions posted online". The Edelman Trust Monitor [13] has similar observations. These are high percentages compared to the 61% trust in advertisements from TV and newspapers, and only 55% in radio. Also, in [20] it is shown that so-called leads that are socially connected are far more likely to buy a product or service when they are influenced by their connections. Part of this is explained by the indirect message being perceived independently and authentically.

All these models and evolutions on social behavior form an interesting basis for interactions on the Internet as well. Many forms of social interactions have been emulated in virtual environments before: forums as public assemblies, expert exchanges such as StackOverflow[1], Yahoo! Answers[2] as expert advice and education, YouTube[3] as entertainment, education and discovery. So when the concept of *social media* was introduced, in the form of Facebook, Hyves, Orkut and Twitter, the intent was clear: create an emulation of the real life social network. Many social networks started out as almost exact digital replicas of the social networks already seen in real life, but soon also made social discovery possible: "befriending" someone who you only know virtually.

It makes a compelling question as to what other forms of social structures are active when communicating through the Internet. Is there a concept of word-of-mouth networks and market mavens? Are there certain people who influence opinion, product sales and concepts more than average?

The characteristics of social media clearly resemble real-world social networks: sharing a messages relates to the spreading of new opinions and news

[1] `http://stackoverflow.com`
[2] `http://answers.yahoo.com`
[3] `http://youtube.com`

with the real-life social network, resharing (or *like/retweet*) is directly related to spreading an existing message/product/opinion, and commenting/replying is related to joining a conversation on an opinion/product/concept in one's social network. It has been found [35] that these modern-day electronic word-of-mouth networks (eWOMs) work similarly as the traditional word-of-mouth networks (WOMs). With the rise of the social media, the word-of-mouth networks that were formerly confined to a certain geographical area, have now been given new boundaries through the internet. What is especially interesting about the online social networks, is that information that is unattainable in traditional WOMs, such as who talks to whom and what they talk about, is publicly available on many social media. Wu et al. [49] found that 46% of TWITTER links reach their recipient not through the original source but through an intermediary. This indicates the existence of extensive word-of-mouth networks, but how do we know who the important "sources" and "intermediaries" are? Who do your customers believe is credible and trustworthy on a brand or topic? Who are the so-called evangelists?

Social media influence has been studied in many different ways. Often used measures are *in-degree* (the number of people following a person), the number of *reshares* or *replies* a message invokes, or a combination to calculate the total number of *impressions* a message produces. Influence has been studied both as a global metric [1], as well as a topic-related issue [8]. Especially the latter study shows that in-degree has a limited correlation with high numbers of reshares. Both studies show that most of the "influential messages" originate from moderately connected persons and not the highly connected elite and content had a high impact on the spread of the message. This implies metrics like Klout [28], which uses metrics like in-degree and number of mentions to decide on influence, are more a popularity score than a true influence metric.

Our hypothesis is that a person's relation to a certain topic is a major factor in his reputation on that topic. He or she might have a high amount of followers that are also interested in the same topic and are more likely to spread the message. We will try to find the attributes about persons that have an impact on both their *reshares* as well as *clicked urls*.

1.1 Defining influence

To determine influence, we should first define what influence is. The Webster Dictionairy [36] defines it as "the power or capacity of causing an effect in indirect or intangible ways". In most marketing or social media papers, this definition is interpreted to mean "generating impressions" or "spreading a message". This seems to stem from the days of broadcasting media, where the number of viewers (impressions) is the key metric for determining influence. In this definition, more equals more. In related research on influence on TWITTER, this idea has been translated to influence measures such as in-degree, retweet

and mention influence. These are measures on how often someone's name has been mentioned in a message.

However, with the data that is available through social media, we might also be able to analyze much more profound metrics. This is something more social media experts are trying to express: stop looking at the number of followers, and start using more profound indications such as Trust, Expertise, Tribes [16]. The only difficulty is that these things are very hard to quantify. Even given a simple definition would be open for arguments. We can agree on what typically would interest a company that is looking for influentials: messages that are turned into sales or other actions that benefit the *bottom line* of the company, be it sales, clients, subscriptions, advertisement revenue, etc. It is not the impressions in and of themselves that matter, but the actions that result from those impressions. In this definition, more may also be less. Arguably, a large number of impressions might lead to a large number of sales, but this might not necessarily be the case.

1.1.1 The sales funnel

In internet marketing, there is a concept of the *sales funnel*. Potential customers or "leads" enter the sales funnel on one side, and paying customers leave the sales funnel on the other side. In between the two are usually a series of steps, such as clicking on a sign-up link, signing up for an account, selecting products or services, entering payment information, executing the payment. The input for this funnel comes from several channels, e.g., advertisements, search engine results and social media. These funnels generally lead to some goal that benefits the company, such as a sale, sign-up, donation, subscription, etc. Ideally, we would measure the effectiveness of an *input* of the sales funnel by calculating the effectiveness of the sales funnel.

If, for example, 2% of the leads that enter the funnel from advertisements proceed to becoming a paying customers, that might be fairly good. However, we only have a single point of reference; what if it turns out that of the leads that enter through social media, over 14% end up becoming paying customers? This would mean that it would be very wise to invest more time on social media, than on advertisements. Our aim is to try to get as close as possible to the end of the sales funnel, to give more accurate knowledge of how social media can increase the exit of the sales funnel. Unfortunately, we cannot measure these sales for they are private information of each company, but we can in some cases measure the number of clicks that lead into the sales funnel by analyzing click data, which is already one step closer than the number of impressions of a tweet, or retweet.

Therefore, our definition of influence is the following:

Definition 1. *Influence in a social network is the ability to generate actions (benefiting the company/topic/subject) of others.*

1.2 Challenges and related work

Trying to find structure and patterns in a large and unstructured network such as the TWITTER graph is a challenging task in the domain of data-mining [27]. This is further explored by the sociological aspect; while in most sciences, the same actions often lead to the same results, in social sciences this is often not the case. The number of factors that play a role is simply too large to take into account. Therefore, even a small correlation between factors can be significant. Also, on many social media, there are factors that are difficult to recognize and filter, such as spammers.

The most interesting work on the challenge of finding influential nodes in large graphs have been from the field of Information Retrieval and Data Mining. A good example would be Page and Brin, who introduced an important influence metric called PAGERANK [31] to their popular search engine GOOGLE. But they lack the distinction of topic-based influence and quantifying the influence to real-world metrics and application. Haveliwala introduced a topic-sensitive PAGERANK metric [19], but this is only applied to TWITTER through an algorithm called TwitterRank [47]. Unfortunately this work lacked motivations on their definition of influence. They assumed the influence of a user is the combination of the influence of his/her neighbors, and the relative amount of content their neighbors receive from him/her. They also use a non-random sample of users, which might cause bias.

With regards to influence measures, Cha et al. [8] empirically investigate the relation between common measures in influence on social media. However, when they test *topical* influence, they only take a small subset of users that have talked about all their defined topics. They find a strong correlation between topics, but in our opinion, this could be caused by their selection bias towards generic TWITTER users, who have a tendency to talk about general topics, instead of also taking into account very topical users that only mention one topic.

1.3 Approach

Our research will be aimed at performing a comprehensive data-mining analysis on topic related influence on online social media. Our main goal is to identify the several types of influentials that have been researched in sociology and marketing. That means not only looking at the most influential people, but also on the less influentials and their role in the spreading of messages. To this purpose, the research will be divided into five distinct steps, each one producing the data needed for the next:

1. In Chapter 2 we will gather a sample dataset from a social network.

2. In Chapter 3 we then distill topical subgraphs from this dataset.

3. Data-mining techniques will be applied in Chapter 4 to find important attributes for evangelism or topical influence.

4. We will then combine these attributes into one or more metrics in Chapter 5 and ...

5. ... in Chapter 5 evaluate this composite attribute.

The steps outlined are schematically shown in Figure 1.1. It shows an example social network, where each node is a user, including information such as the content of sent messages. These users are subjected to both graph analysis as content analysis, dividing the graph into topical subsets, in which each user may be more or less active (shown as **boldness** of the node). These will generate the data we need to train a classifier on certain *ground truths* of influence, which will be defined in Chapter 4.

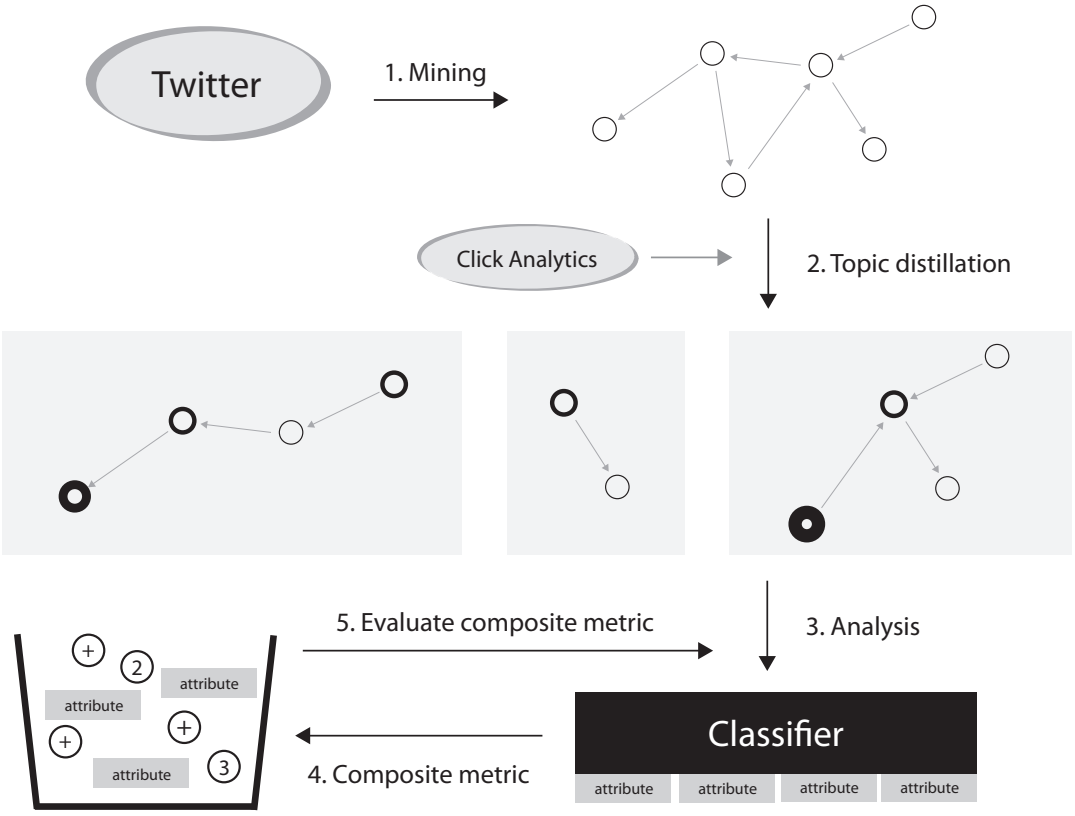

Figure 1.1: Schematic representation of the approach of the research.

Our approach is aimed to use topical influence on brands, therefore we want to have control on the definition of the topics. Our influence measure will reflect this also; we will define influence as close as possible to the sales process described in Section 1.1.1.

Technically, we are limited by the size of online social networks, and will use sampling of the social network of choice, TWITTER, to get a representative subset. We compare characteristics of this subset to datasets used in similar research, to verify the similarity of the subset. We also use platform specific features of the social network TWITTER, mostly based on hyperlink redirection, that makes this research hard to replicate on other networks. Also, certain design decisions of TWITTER are particularly useful in forming topical relationships, as will be further discussed in Chapter 2. Especially the amount of *public* communication on TWITTER creates an unique opportunity for this type of research.

Because of our feature construction approach, we have implicitly limited the *origin* of influence to certain factors, such as the popularity of a user and their use of a topic. This approach is susceptible to missing factors that may determine our definition of influence, in favor of creating an easily interpretable explanation of influence.

Chapter 2

Mining the social web

First, we will provide a short introduction to the social network of choice, Twitter. Then, we will describe our method for collecting a sample graph from the network.

2.1 Introduction to Twitter

Twitter[1] is an online social network, which is characterized as a social *graph* where users are interconnected through relationships and interactions, and are able to share information among each other. It was founded in 2006 and its core features are the ability for Twitter users to post messages (*tweets*) with a maximum length of 140 characters to their profile, and the ability to *follow* other users. The tweets of these other users then show up on their personal *timeline*. Relations on Twitter are unidirectional, creating a directed social graph; being connected to a user, does not automatically mean that the user is also connected to you. This is contrary to the bidirectional connections of other popular social networks such as Facebook, Orkut, Hyves and MySpace.

In their messages, Twitter users can reference (*mention*) each other by prepending an `@` to the referenced username. A special reference is called the *retweet*, which is a mention prepended by `RT` and appended by a exact copy of the original content of the tweet. This method is used to share content and propagate messages through the social network. These types of messages have often been used in research as influence measures [], with the reasoning that being mentioned a lot, or having one's message spread through many users is a sign of being either very popular of very influential.

We will be using Twitter as our social network for data-mining for several reasons. As mentioned before, in many other social networks, connections in Twitter are directed, and thus not necessarily reciprocal. This feature affects many other aspects of Twitter. For example, it means that users can engage with people they are *topically* interested in, but are not real-life *friends* with (such as celebrities, political figures, industry leaders, etc.). That is, the

[1] `http://www.twitter.com`

content of the messages is the primary cause of the relation, rather than the existing real-life social connections. This unique effect makes the TWITTER network interesting for this specific research; more than only a personal network, content might play a large role in whether someone is popular and/or influential. Research [8] has shown that reciprocity on TWITTER is low ($\sim 10\%$), suggesting that the network is largely based on one-way 'interest' relations. We get two important measures from this directed social graph: the *in-degree* is the number of people following a user, while *out-degree* is the number of people a TWITTER user is following him/herself (the *friends*). The in-degree has often been used as a popularity and influence measure. This is based on the reasoning that with a high in-degree, one's messages are being read by a large group of people (a large *audience*), giving a high number of impressions, and may thus impact the decisions or opinions of many other people.

Also, because it so easy to gather a group of followers when your messages are public, a much larger percentage of TWITTER users have a *public profile*, allowing one to track many of the conversations that take place on the network. While collecting our data, we found that only 20% of TWITTER users have protected their messages from public access. A recent survey [34] mentions the average of protected profiles on online social networks is 58%.

2.2 An example

We will first introduce an example of a (very) small graph, that we will use as a running example in the subsequent chapters to explain our methods. The graph in Figure 2.1 consists of 5 nodes. We show the "following" relation with the edges of the nodes, and the content produced by these fictional users is shown in the boxes in Figure 2.1. We will reference this example a number of times in this research to explain the workings of topic distillation and a number of analytics.

2.3 Accessing the network

Collecting messages from TWITTER can be done in several ways. First, there is a resource named "the Firehose" [46]. This is a stream of *all* of the messages sent through TWITTER, which as of November 1st 2011 is about 200 million per day[2]. This stream reports on all content and users, but not on the social graph itself. For obvious reasons, this stream is considered a very valuable resource and access has been limited since TWITTER's early days. Resellers are available, but are still costly for a research project like ours.

There is a pubicly available sample stream [46], which streams a random subset of tweets, estimated to be around 1% of the total messages. There is also a request-based REST interface [45] that supplies information about profiles,

[2] `https://dev.twitter.com/discussions/3914`

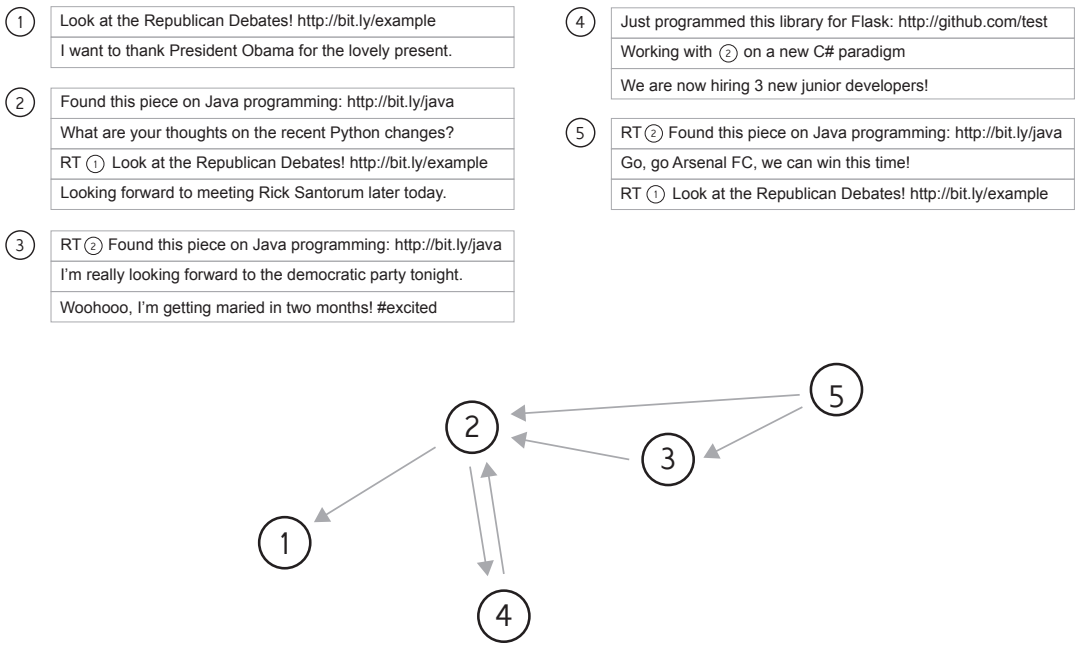

Figure 2.1: Example: a fictitious Twitter-like social network.

tweets by specific users, friend/follower information, etc., which is limited to 350 requests per hour. Contrary to the streaming methods, this interface also provides information about the social graph. One may request the connections of others, as well as their posted messages and retweets, provided they are public messages.

In experiments to get a reasonably deep (rich user information, accurate knowledge of the social graph, plenty of tweets per user) subsample of the Twitter social graph, using the sample stream was unsuccessful; most of the messages that were received, were from different users. When trying to enrich the stream by collecting graph connections through the REST interface, the request rate limit caused not only a major time cost to complete the dataset, but also implicitly limited the dataset to the set of most active users.

Instead we will be solely using the Twitter REST interface. We crawled a subset of Twitter users using a Forest Fire algorithm, which was proven to be the most effective means to sample a large graph in [32]. The steps are outlined in Section 2.4.

2.4 Graph sampling

Our sampling of the Twitter graph was performed using the social network graph first, and only gathering information about content second. The sampling described in this section therefore, has no relation to messages or content of messages.

Because of the limitation set by the request rate limit, it will not be possible

for us to gather a complete dataset of the entire TWITTER network. Therefore, we will have to retrieve a representative random subset of the original graph. Leskovec and Faloutsos [32] found that a Forest Fire (FF) algorithm was the most representative way to sample from large graphs in general. Forest Fire is an algorithm that picks a random node of the graph and starts randomly "burning" the outgoing edges recursively until the "fire" is stopped, when the probability decides no more edges should be burnt (akin to a series of consecutive heads when flipping a coin). It then picks another random node and repeats the process. Our sampling algorithm will be a FF algorithm with a forward burning probability $p_f = 0.6$, which was found to be a good value for our kind of graph in the aforementioned paper [32]. We will then compare the characteristics with some other papers in Secton 2.5 to see if our dataset has corresponding properties. The crawler ran three times, with slightly adjusted parameters, collecting up to 33,000 twitterers. This should provide decent samples of the TWITTER graph to use for subsequent chapters. The steps we are using to gather the subsample are:

1. Select a TWITTER user i (by randomly selecting a TWITTER user ID).
2. Retrieve all of i's friends connections and store them into an adjacency list A_i.
3. Select p random friends from A_i (p following a geometrically random distribution), forming a subset $X \subseteq A_i$ with $|X| = p$.
4. Repeat steps 2 and 3 for each $j \in X$.
5. If there are no more nodes to visit, start at 1.

To get to our aimed goal of gathering *topical* and thus contextual influence on the network, we also collected up to 1,200 of the most recent messages from all visited TWITTER users. These messages will be used for topic distillation in Chapter 3 to provide context for their influence performance.

2.5 Dataset characterics

Before we use the data, we will first look at some of the characteristics of the datasets. We ran several graph metrics on each of our datasets, which only differ in the length of time they ran, and thus the size of the crawl. We labeled these datasets *small*, *medium*, and *large*.

We further analyze the graph with common graph metrics. Before online social media existed, most of these metrics were already used to investigate real-life social networks [42]. Each metric quantifies some property of the graph, and tries to explain the structure of the graph. The metrics investigate the way the

nodes are connected, and if there are common patterns that can be identified. The results can be found in Table 2.1, but we will first explain these metrics:

Average degree is simply the average number of outgoing and incoming edges of each node. Put simply, it equals N_e/N_n, where N_e is the number of edges in the graph, and N_n is the number of nodes.

Density represents the density by comparing the number of edges N_e in the graph to that in a *complete* graph with the same N_n (i.e., where all nodes are connected with all other nodes). Formally, this metric is defined as: $N_e/(N_n * (N_n - 1))$.

Modularity is a metric that tries to decompose the graph into *modules*, in which nodes are highly-connected, but less connected to nodes in other modules. It achieves this by creating possible clusters in the graph and measuring the density inside the cluster and comparing it to the density between clusters. If the density from one node to nodes in a cluster is high, and the density from the same node to nodes in other clusters is low, modularity is high. In the algorithm we used, the clusters were assigned using an algorithm proposed by Blondel et al. [6]. For the current research, this will indicate if there are either many groups of friends or topically-related people (which would be of particular interest to us), or little; which would indicate that the social network does not contain groups.

Average Local Clustering Coefficient ($\overline{cc}$) measures the degree of clustering in the graph. It looks at the individual nodes of the graph and its neighbors and compares the connectedness with a *complete* graph. A high value for this metric is known to show the "small-world" effect and is an indication of the way nodes are connected with their neighbors. So, while **Density** and $\overline{cc}$ are both measures of graph density, $\overline{cc}$ measures the density on a very local level, while **Density** primarily is an indication of global density.

Diameter is the length of a longest path between any two nodes in the graph.

Average Path Length ($\overline{d}$) is the average path length between two nodes. This metric is expected to comply with Milgram's famous 'six degree of separation' experiment [37]. Our average path length is very close to the values found by Kwak et al. [29] in their quantitative Twitter research; they found an average path length of 4.12.

Hyperlinks to external webpages are found in 34% of tweets. The average number of tweets that contain mentions of other users is around 53%. The number of retweets is around 32% of the total tweets. Of the retweets, 33% contain at least one link to an external resource.

Reciprocity, meaning bi-directional connections, thus following each other, was found to be 13.65% of all the connections in the graph in the medium

dataset, which is the same order of magnitude as the numbers found in [8, 29], where a near-complete graph of TWITTER was used.

We have developed the methods used in further chapters on both the small and medium datasets, to keep runtime to a minimum during development, but all the results in this research paper (unless otherwise mentioned) are gathered using the large dataset as source. This is the most representative graph and provides us with more data to work with.

	small	medium	large
Nodes	1,832	8,396	31,891
Edges	6,543	85,350	584,661
Degree	3.581	10.166	18.333
Modularity	0.474	0.416	0.471
Density	0.002	0.001	0.001
$\overline{cc}$	0.092	0.114	0.068
Diameter	15	15	13
$\overline{d}$	4.778	4.182	4.027
Sample date	6/01/12 – 9/01/12	16/01/12 – 31/01/12	31/01/12 – 27/02/12

Table 2.1: Dataset characteristics

2.6 Empirical analysis

What is immediately apparent from looking at the large dataset in Figure 2.2, which was organized using Gephi [2] and its Force Atlas algorithm, is the existence of a few very well-connected users. They have a very large in-degree, network centrality and high PAGERANK [31]. However, the in-degree distribution follows a very sharp decline. As can be seen from Figure 2.3, the in-degree frequency is decreasing exponentially. While the user with the highest in-degree of the sample has 2,855 incoming edges, the average is only a little over 18. This is in accordance with previous work [1, 47, 8].

Next, we decided to look at a list of high-profile users. We have ranked the TWITTER users in the sample by their PAGERANK and the ten highest ranking users are shown in Table 2.2. Users are shown with their in-degree (number of followers), out-degree (number of friends), eigenvector centrality (a measure of influence of the user in the entire network) and the local clustering coefficient (a measure of how close the users' neighbors are to forming a *clique* [38]). It is clear that the top users in our sample are also top users in real life. To compare our subsample with the full graph, we also included the number of followers for each of the ten users.

After further inspection by visualizing the graph (again using Gephi's Force Atlas algorithm), the TWITTER users form clusters that seem to have a topical relation. Sport players are clustered with each other, as are politicians, as are technology blogs/influencers, as are social media gurus, etc. This strengthens our beliefs in the hypothesis that there is some topical relation at play in the forming of these social networks. This effect can be seen from the overview of the social network in Figure 2.2.

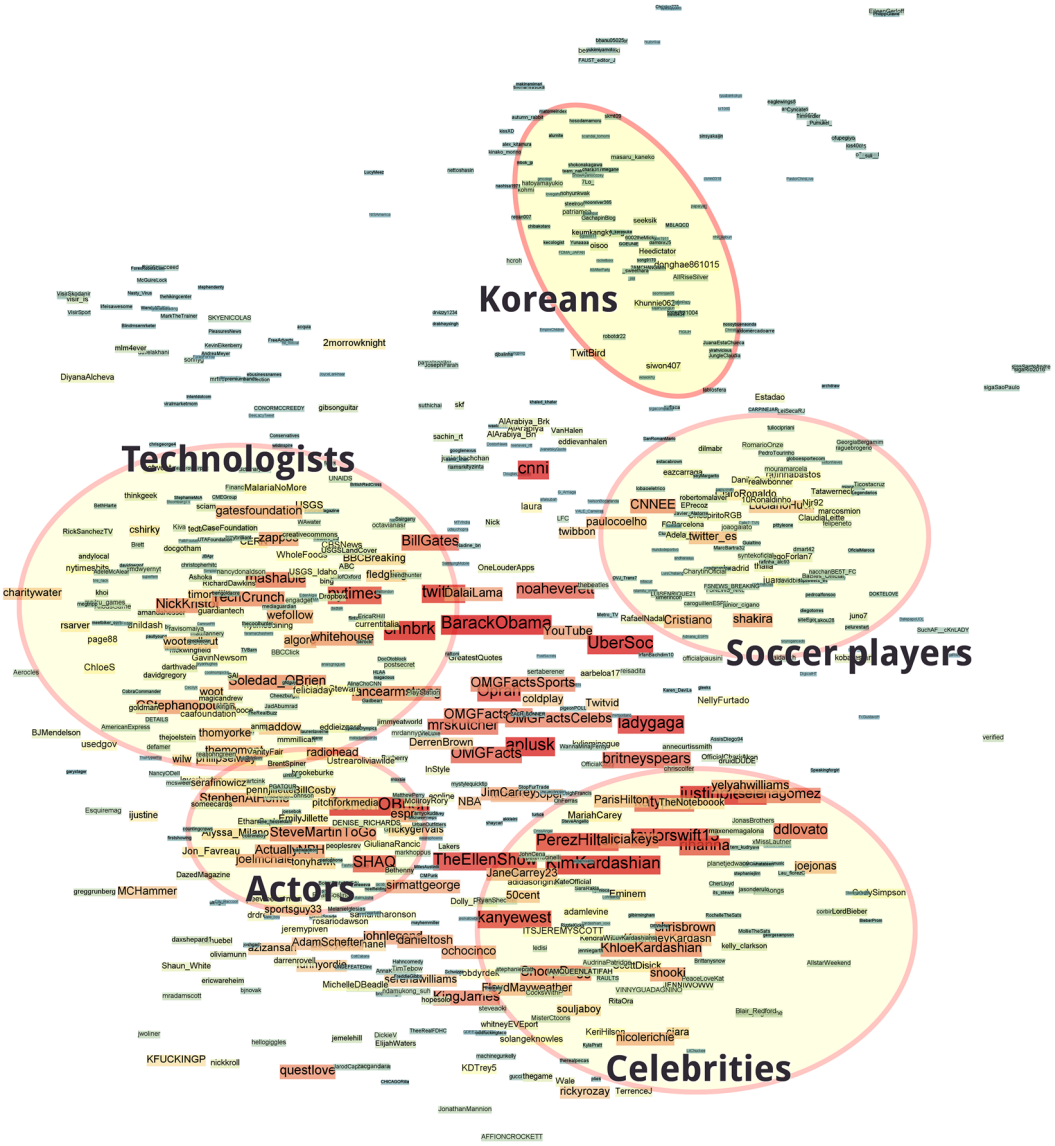

Figure 2.2: Manually annotated visible clusters.

Screen Name	PageRank	In	Out	Eigenvector	Cluster	Followers[a]
BarackObama	0.0043	2,855	668	1.0	0.0062	12,966,898
ladygaga	0.0024	2,501	88	0.7182	0.0075	20,612,266
justinbieber	0.0024	1,974	404	0.6439	0.0119	18,357,753
katyperry	0.0014	1,848	30	0.5271	0.0106	15,997,487
Oprah	0.0024	1,751	21	0.7075	0.0156	9,768,533
TwitPic	0.0025	1,731	83	0.4725	0.0	6,600,021
KimKardashian	0.0020	1,712	63	0.5601	0.0130	13,882,950
britneyspears	0.0013	1,694	271	0.5380	0.0098	13,942,698
aplusk	0.0024	1,689	177	0.7163	0.0179	9,751,009
TheEllenShow	0.0024	1,683	1,348	0.7092	0.0164	9,869,979

[a] Snapshot from 12 March 2010

Table 2.2: Top users using global metrics, ordered by in-degree.

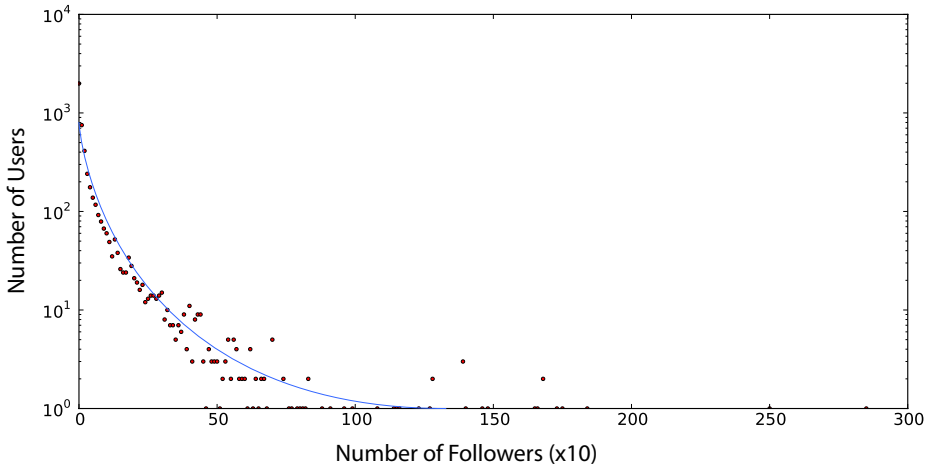

Figure 2.3: Distribution of in-degree: number of users with a certain number of followers, on a logarithmic vertical axis.

2.7 URL analytics

As can be seen in Section 2.5, almost 34% of the messages on TWITTER contain a link. Because we are interested in the *content* of the messages, we can attempt to use these links to get some additional data about the user who has sent them, and the impact of the particular message. In this section we will explore the use of links on TWITTER and discuss the possibilities of using them as additional information in our analysis.

First and foremost, we might attempt to collect analytics on these links and see if any influence can be derived from the resulting data. For example, if a user shares a link which is seen by other TWITTER users 1,000,000 times and clicked on 100 times, while the same link shared by another TWITTER user is only seen 100,000 times but clicked 2,000 times, we might argue that since clicks are further down the *sales funnel* than views, the latter user is actually *more* influential. Specifically, if we build our topics around brands, this might be directly related to our definition of influence: generating more clicks for a certain brand will cause input into the sales funnel and thus influence sales.

For proper analytics of the links we need to know the answer to the question: "How many clicks are originated from a given tweet?". In most circumstances, this kind of data can only reliably be obtained through analysis of clicks on the individual websites the links refer to. This information is only available to the website owners, and not publicly available. However, this is where the 140 character limit of TWITTER is of good use.

In most cases on TWITTER, people wanted to share large links, and comment with text in the same message (e.g., "Look at what I just found! `http://www.example.com/blog/2010/6/12/look-at-what-I-found`"). However, because of the character limit a type of service called the *url shortener service* was invented. This type of service takes a large link, and uses an algorithm to generate a unique, short url, which can easily fit in the 140 character limit (e.g., `http://sho.rt/a4bCz1`). This provides us with another layer to inspect analytics on. We found one URL shortener that supports publicly available analytics at all, which is `bit.ly`[3]. For every shortened link, we can request the number of clicks coming from a different *referrer*. In the past, TWITTER click analytics has always been tricky, because different tweets were often displayed as the same referrer in the analytics. However, in early 2011, TWITTER introduced their own URL shortener `t.co`, which shortens *all URLs*, even those that have already been shortened by `bit.ly`, and redirects them to their proper destination. By introducing this extra step, which sets the referrer to the unique `t.co` URL, we know for certain that analytics are originated from a unique tweet, or its native retweets.

In our experiments, we found 21.3% of tweets containing links (34% of the total number of tweets), had their URL shortened by `t.co`. Of that, 33.2% was

[3] `http://bit.ly`

also shortened by `bit.ly`. These are the tweets that we can properly perform click analytics on. This is not a large amount of tweets, but given the size of the large dataset, the number of links will still be reliable. When using a larger sample, or even a complete dataset of Twitter, the results will become even more reliable.

Furthermore, click analytics may be able to provide a guard against spammers and bots, because unfortunately, there are many bots and spammers active on Twitter. One way they oftentimes try to gather a following is by automatically retweeting messages from popular Twitter users. These retweets however do not generate any valuable impressions or actions and only promote the already popular users. Clicks however are not likely to be emulated by bots, so this could be useful in circumventing bias.

Chapter 3

Topic distillation

In this chapter we will discuss the possibility of generating content related attributes for our dataset. We have collected up to 1,200 tweets for every user in the graph that we can use to analyze this information and use the content of these messages for our analysis in Chapter 4.

3.1 Distillation techniques

Topic distillation, also called concept mining, subject analysis, topic discovery, or topic modeling, is a field of research where the goal is to extract as concise and brief information as possible from a perhaps large dataset [4]. In our case, the subjects are relatively small 140-character messages, but in many other text classification applications the contents might be much larger. Search engines, for example, use these techniques to condense web documents to a set of topics and return search results that are more accurate than simple keyword-based matching.

In general, topic distillation techniques will define *topics*, consisting either of n-grams (sequences of n words or characters) or a vector of keywords. The definition of the topics can be done manually or using an automated algorithm. While the former requires knowledge and understanding of the content of the corpus, it does lead to more accurate results if the topics are defined properly, while the latter technique can be used on any dataset.

Common algorithms for topic distillation are probabilistic Latent Semantic Analysis (pLSA) [12] and Latent Dirichlet Allocation (LDA) [5], which have a very similar theoretical basis. We consider Latent Dirichlet Allocation, which is a generative model that assumes documents are a mixture of a number of topics, and each word in the document is generated from a topic. The *topics* it produces are multinomial distributions over words that could be generated by that topic. First, the algorithm considers all messages as *bags of words*, meaning the sequence of the words does not matter; only the frequency of occurrence matters. Then, the algorithm starts with a random distribution of the words in the corpus to K topics. An inference algorithm is used to

train the topics to the documents in the corpus. Most commonly, this inference algorithm is a Gibbs sampling algorithm [14], a randomized algorithm used to approximate probability distributions. A topic resulting from the inference might for example be characterized by the following distributions: 50% cat, 20% cute, 10% horse, 20% adorable. Its perfect message would contain words exactly in those proportions (e.g., "the adorable horse and cat followed the cute cat to the cat that was hugging the adorable cat. cats are cute."). This topic could be interpreted to be about animals, although the algorithm is not context-aware so this interpretation is quite difficult in some situations.

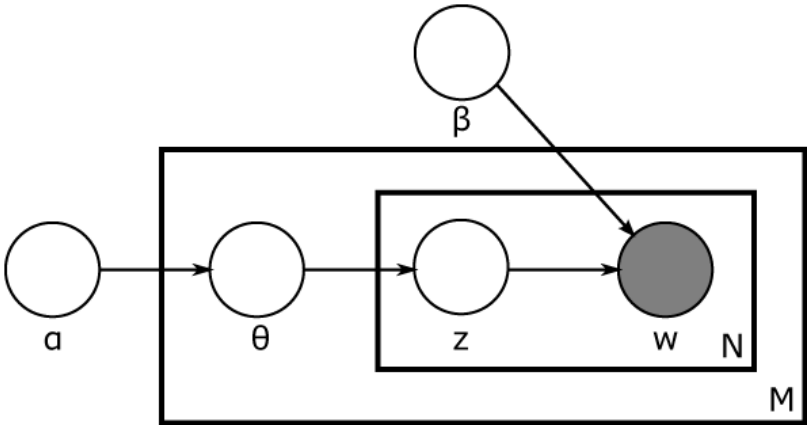

Figure 3.1: Schematic notation of the LDA model. The outer rectangle represents documents and the inner rectangle represents the repeated choice of topics and words within a document.

The probalities of the LDA model as shown in Figure 3.1 are defined by:

α is the parameter of the Dirichlet prior on the per-document topic distributions.
β is the parameter of the Dirichlet prior on the per-topic word distribution.
θ_i is the topic distribution for document i,
z_{ij} is the topic for the jth word in document i, and
w_{ij} is the specific word.

3.2 Topics on Twitter

Regardless of how we define and extract the topics, the result of the topic distillation will consist of one or more attributes about the messages from Twitter users. These are our *topical attributes* which we will use to analyse the impact of content in the influence of Twitter users. We can condense the attributes on messages to the Twitter users themselves, so we get the frequency of usage of the topic, the ratio of their messages on this topic compared with their total number of messages, etc.

To test our methods, we also used a widely utilized citation dataset, of which the abstracts of the papers will simulate the content of tweets, and authors of papers represent TWITTER users. Since this dataset is a complete collection, in contrary to our sampled dataset of TWITTER, it should help us to increase the reliability of the theories described in this chapter.

Influence is explicitly defined on the entities of authors/persons. So in the topic distillation, we will have to detect and reduce topics from the papers and tweets (*documents*) to their respective authors. One can visualize this process as creating a one-to-many relation between author and his/her documents, and creating a one-to-many relation between between documents and topics. Lastly, the topic information is aggregated to the author, so an author is now associated with a list of topics.

3.3 Citation dataset

To experiment with how to properly extract topics from a graph with meta-information, and to prevent potential errors in the TWITTER dataset to interfere with the topic distillation, we have first tested these methods for topic distillation on the HEPTH citation dataset [33]. This dataset offers a full history of all scientific citations in the field of high energy physics for the period 1992–2003, including some paper meta data like title, authors, etc. When we compare the kind of data we are looking for, it is not difficult to see the similarities with the TWITTER dataset: authors (persons) write papers (tweets) and in doing so, reference (mention/retweet) other authors (persons).

Using a measure of importance, namely the subjective importance of an author in any particular field, we experimented with several algorithms to generate attributes on the authors, such as PAGERANK, HITS, betweenness, etc., and primarily used these as lessons for the TWITTER dataset. Due to inaccuracies in names of authors and the use of institutions in the names of authors, the results were not as reliable as we hoped and we did further experiments on the TWITTER dataset. We did, however, find that the HITS and PAGERANK algorithms produced promising results in ranking popular users.

3.4 Twitter dataset

In most of the papers on content-based analysis on TWITTER, either Latent Dirichlet Allocation (LDA) or keyword matching is used [8, 1, 47]. However, in TwitterRank [47], the use of LDA causes some debatable results. The authors of [47] generated topics contained many of the same keywords (which makes those keywords irrelevant) and the topics were very difficult to interpret. Arguably they failed to serve their purpose of topics altogether, but at least their results produced topics that we, in this research, would not find useful.

When we tried to use LDA for topic distillation, we found similarly confusing

topics. An overview of the most significant topics and the related words, in order of descending probability, can be found in Table 3.1. However, our goal in this research is to investigate the use of topics that are comparable to the use of brands, interests and fields. These topics do not have to be *exhaustive*, but there has to be a clear field of interest and a certain sociological market/group for each of the topics. A more precise (in the sense that a whole field/brand is covered in the topic) is of course preferable, but the size of the data would still allow us to see the same patterns we would expect to see using a less accurate description of a topic.

topic	keywords
#1	twitter, boy, twitpic, pretty, haha, nice, miss, hahaha, game
#2	time, thing, ff, guy, back, people, make, question, today, gonna, girl
#3	blog, art, post, design, climate, car, twitter, top, flu, recovery
#4	day, radio, green, card, dream, sound, food, san, heart, bank, chart, car, coffee, drink
#5	vote, today, show, god, sign, winner, join, dog, congrats, day, family, free, brown, wow, omg
#6	lol, man, van, haha, justin, met, shit, f*ck, dat, b*tch, n*gga, lmao

Table 3.1: Results of Latent Dirichlet allocation.

Therefore we have decided to create topics based on keywords inferred from Wikipedia articles. Other papers that have used keyword-based topics [8, 1] are often using a feature of Twitter, called *hashtags*: any term preceded by a hashtag (#) is linked to other messages that contain the same hashtag. Compared to this approach, our method is more generic and does not rely on the use of hashtags, while it still includes hashtags. i.e., if one of our keywords is "network", then the word "network" used in a sentence, as well as the hashtag "#network" will be matched. Bakshy et al. [1] use empirically selected topics by manually binning messages into topics, which might be even more reliable since there is no confusion about context, spelling, etc. Tweets are *on-topic* when they contain at least one of our selected keywords. A tweet might be on-topic for multiple topics.

Our topics should be non-overlapping and clearly have different types of people interested in them. By inspecting the timeframe of the messages in the dataset, one can make reasonable topics based on world events that occurred, but also on topics that are mentioned constantly on a social network such as Twitter (e.g., celebrity gossip, programming, fun facts, etc.). By inspecting

Topic	Keywords
Politics	democratic republican democrats presidential political election republicans government executive constitution federal senators elections congressional representatives elected politics presidents obama biden gingrich perry romney santorum
Tech	web internet www html computer data software online browser oss opensource "open source" programmer programming developer code coding java c c# c++ php "visual basic" python objective-c perl javascript sql ruby haskell perl actionscript
Obama	obama
Premier League	arsenal blackburn chelsea liverpool ... *(list of all clubs currently in premier league)*

Table 3.2: Keywords of the predefined topics.

the types of users, starting for example with the dataset overview in Figure 2.2, one can also find some very different use cases for using Twitter, be it gossiping, talking about fashion, programming, world news or sports.

3.4.1 Wikipedia ontology

Of course, using predefined topics can cause bias towards being either too specialized or too generic, giving an advantage to either very topical people, or the very generic ones. To try and circumvent this and make sure we have relatively complete and reliable topics, we will base our topics on the most frequently used words in predefined Wikipedia [48] articles. We have manually selected a few topics that instinctively have little contextual overlap. We then removed ambiguous words, that could be interpreted differently from our intended context and meaning, such as "foot", which may mean a physical foot, attached to one's leg, or a unit of length. Our resulting topics can be found in Table 3.2.

Additionally, to the "Politics" topic, we appended the last names of the current presidential candidates and the names of the current president and vice-president. Also, to the "Tech" topic, we appended the most popular programming languages as listed on IEEE's blog [25].

In Figure 3.2 one can see how this topic distillation works on our example, for the topics specified in Table 3.2. **Bold** words are part of the "Politics" topic, *italic* words are part of the "Tech" topic and underlined words are part of the "Premier League" topic.

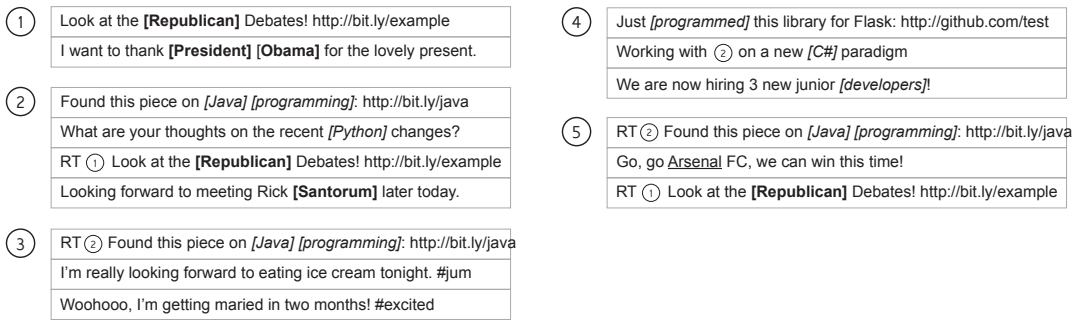

Figure 3.2: Topic distillation in our example.

3.4.2 Resulting topic graphs

In this section we will describe the topic graphs that result from the topical analysis, and which will be used for influence analysis in later chapters. We will first look at some of the newly introduced attributes that the topic distillation has generated for our dataset. The most important of these is the *topical ratio*: the ratio of messages on-topic on the total original tweets of the user. This metric is also a decreasing power-law function, as can be seen from Figure 3.3.

We only included a user in the topic graph if he or she mentioned the topic in at least 0.5% of his/her tweets; the topical ratio must be at least 0.005. Given the maximum number of collected tweets (1,200), the minimum number of on-topic messages must thus be greater than $1{,}200 * 0.005 = 6$ in the vast majority of users. This eliminates accidental inclusion of one-time topic participators.

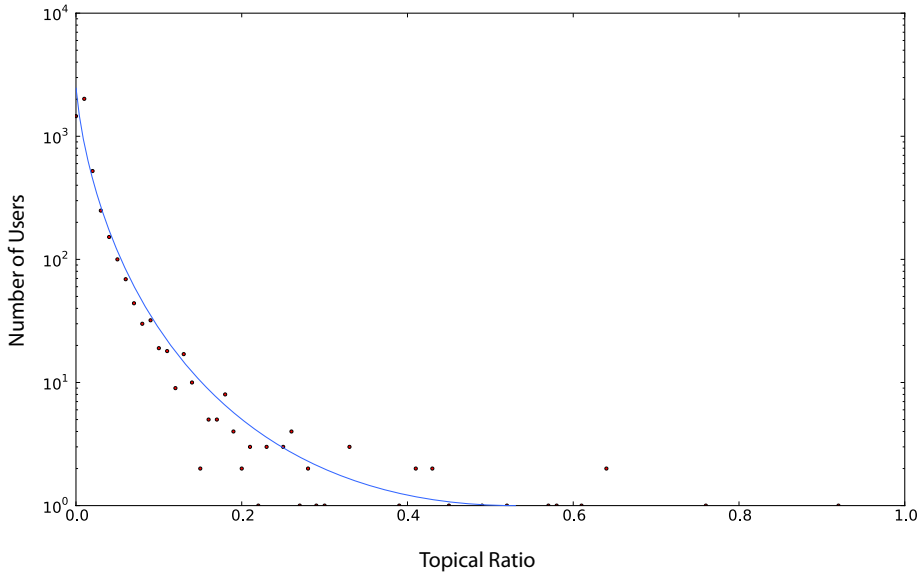

Figure 3.3: Histogram of topical ratio, vertical axis logarithmic.

Now we will also compare the metrics of these new topic graphs with the metrics of the global graph (found in Table 2.1). The results of this comparison

can be seen in Table 3.3. When comparing the metrics, we noticed a few important differences. It seems that the more specialized the topic is, the larger the clustering coefficient and the smaller the modularity. This can indicate that there are less sub clusters within the network, and the amount of clustering withing the network is larger. We would expect to see these metrics increasing if we took a sample from the graph that is *more connected* than a random subset. This can indicate that the more specialized (less keywords or a specialized subject) the topic is, the more clustered together the network is, while degree often stays similar.

One might notice that the "Premier League" topic is a special case: in all aspects it looks more like the global graph (several clusters/highly modular, not very dense, etc.), but it has a higher clustering coefficient. This seems to contradict the statements in the previous paragraph, but we think this may indicate that the people interested in their premier league club might cluster together, but they do not follow many others from other (competing) clubs.

For each of the topic graphs, we link all content information based on what is applicable to the topic, so that only topic-related messages are attached to the topic graphs. This includes messages sent by the user, retweets and mentions sent and received by the user and also URL analytics (see Section 2.7) for each of the URLs posted by this user on this topic. In this way, the identification of these subgraphs also prevents overfitting of solutions in Chapter 4 because we gather our goal attributes on a per-topic basis. It is therefore very unlikely that a trend that is common over two or more of these topic graphs is due to a too specialized solution, thereby overfitting the classifier.

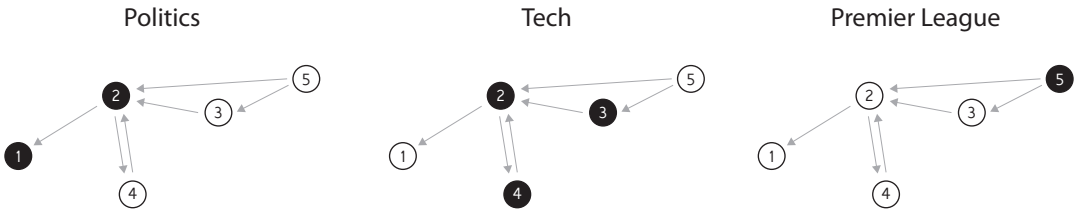

Figure 3.4: Topic graphs in our example.

Concluding, from the global graph, we have extracted several topic graphs using the methods described in this chapter. Figure 3.4 shows the resulting subgraphs when applied to our example (the semitransparent nodes are removed from the topic graphs). Each node in the topic graphs has meta-data about the content they sent and received. This graph and meta-data will be used to construct features and analyze these features for topical influence in the next chapter.

	Original	Tech	Politics	Obama	Premier League
Nodes	31,891	3,109	1,815	816	897
Edges	584,661	72,213	41,678	15,609	8,057
In-Degree	18.333	23.227	22.963	19.129	8.982
Modulariry	0.471	0.330	0.286	0.234	0.409
Density	0.001	0.007	0.013	0.023	0.010
Avg CC	0.068	0.164	0.184	0.204	0.162
Diameter	13	11	11	10	12
Path Length	4.027	3.273	3.113	2.904	3.965

Table 3.3: Topic graph characteristics.

Chapter 4

Analyzing influence

Starting from the theoretical definition of influence in Chapter 1, we will now create some very specific practical influence definitions we can use to analyze the topic graphs that we have extracted in the previous chapter. First, let us repeat the original theoretical definition given in Chapter 1: "Influence in a social network is the ability to generate actions (benefiting the company/topic/subject) of others".

4.1 Definition

Within our dataset there are several attributes of every TWITTER user that we can interpret as "an action generated by others". The most often used measures in social media research fits well in this description, because a TWITTER user mentioning or retweeting another TWITTER user can be defined as "an action generated by others". In-degree influence can be seen as generating views of the message. And another common measure of *audience* influence, is a combination of the two former definitions by enlarging the number of views by generating retweet actions. Additionally, we will introduce a measure of influence in social media that to our knowledge has not been used before: the ability to make other TWITTER users click on a posted link is also a measure of influence.

Because in-degree influence has been studied a lot already, we will try to explain influence using two different possible definitions.

Definition 2. *Influence on* TWITTER *is the ability to generate clicks on posted URLs.*

Definition 3. *Influence on* TWITTER *is the ability to generate retweets on posted messages.*

The results from the attribute analysis in this chapter will be used to experiment with *combinations* of attributes. These combinations will then also be analyzed using the correlations and data-mining techniques found in this chapter, to create a feedback loop (recall Figure 1.1) that will define a correlated

metric for influence, which can be used as a predictor for topical citations/tweets for any user.

4.2 Attributes

Using the topical graphs as described in Chapter 3, we will now expose several attributes of TWITTER persons and use data-mining techniques to detect patterns between the attributes and *the ground truth(s)*: features that represent our definition of influence, if there even is such a pattern. This provides us with information to determine the major components of our definition of influence on TWITTER. For each of these attributes we will indicate the complexity of retrieving the attributes when we need to build the dataset. Keeping complexity low for the composite prediction attribute(s) is important, because the faster we can calculate the new metric, the larger the dataset of TWITTER users we can populate and thus the more reliable our predictions can be. In these complexity classes we will use n for the number of TWITTER users, m for the number of followers of a TWITTER user and p for the number of messages from a TWITTER user x.

Note that these complexities are reflecting very crude and brute-force methods, and in practice several methods may be combined and optimized when taken into practice. Nevertheless, it is a good indication and because of our limited sample size, the required computational resources remained very reasonable, even on the most complex attributes. Furthermore, we will use the following attributes in the explanation of the attributes:

O_x The set of nodes connected to x through x's outlinks.

I_x The set of nodes connected to x through x's inlinks.

M_x The set of all messages sent by user x.

M_{tx} The set of all messages on topic t by user x.

C_{tx} The set of all links on topic t by user x.

R_m The set of all retweets of message m.

G The set of nodes in the "large" global graph.

T_t The set of nodes in the topic graph of topic t.

4.2.1 Followers and friends

First and most predictably, we will use the number of followers (in-degree) and number of friends (out-degree). It is important to note that these numbers are a snapshot, taken while collecting the dataset. A more complete and correct metric would be to have an average of the number of followers over a certain

time, but this is only possible when measuring followers at the time a message is sent; this would take time but is possible using the streaming methods of gathering data. Complexity of computation for these attributes is $O(n)$. We generated the following attributes:

1. Number of *followers*: The total number of people that have subscribed to the user's messages in our subsample of the graph. We will denote this by $|I_x|$.
2. Number of *friends*: The total number of people that this user has subscribed to, denoted by $|O_x|$.

4.2.2 Mentions and retweets

Traditionally, these metrics have been an important part in the research as *targets* of influence measures. We show these metrics on our example in Figure 4.1. Other research has often used these metrics in the sense that being talked about is an important aspect of being influential. We will first use these attributes as source attributes, and later use them as target attributes instead. Complexity of computation for these attributes is $O(n * p)$. We propose the following attributes:

3. Total topical mentions: $m(t,x) = \sum_{i \in T_t} |\{j \in M_{ti} : x \in mentions(j)\}|$, where $mentions(j)$ is a function that extracts the set of users mentioned in messages j.
4. Total global mentions: $m(x) = \sum_{i \in G} |\{j \in M_i : x \in mentions(j)\}|$, where $mentions(j)$ is a function that extracts the set of users mentioned in messages j.
5. Mean topical retweets per message: $\overline{rt}(t,x) = \frac{1}{|M_{tx}|} * \sum_{m \in M_{tx}} |R_m|$.
6. Mean global retweets per message: $\overline{rt}(x) = \frac{1}{|M_x|} * \sum_{m \in M_x} |R_m|$.
7. Mean topical retweets per message, per 1000 followers: $rpm(t,x) = \overline{rt}(t,x) / \frac{|I_x|}{1000}$.

4.2.3 PageRank

PageRank [31] is a link analysis metric on graphs that can be interpreted to determine direct and indirect influence of a graph's connectivity. It is most famously used in the Google search engine to determine relative importance

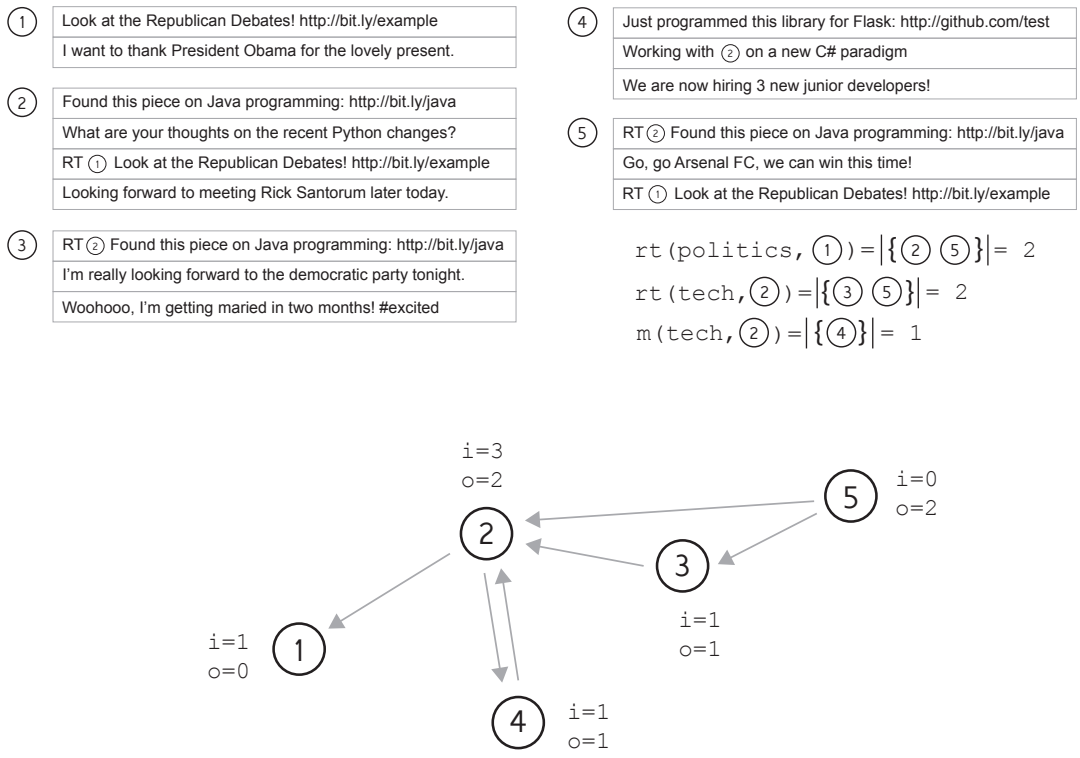

Figure 4.1: Followers ($|I_x|$), friends ($|O_x|$), retweets (rt) and mentions (m) in our example.

of web pages by looking at the way other pages link to it. It uses the instinctive reasoning of a *random surfer model*. This model simulates a surfer that selects a random node, and repeatedly keeps following a random outlink with probability d, or jumps to a random node in the graph with probability $1 - d$. This probability d is called the *dampening factor* of the algorithm and prevents pages that constantly only link to each other (many times) to cause skewed scores.

We denote the PageRank of a user x by $PR(x)$. In the most simple form, given a node x, with incoming links (in our case, *followers*) from nodes a, b and c, we have

$$PR(x) = \frac{(1-d)}{n} + d\left(\frac{PR(a)}{L(a)} + \frac{PR(b)}{L(b)} + \frac{PR(c)}{L(c)}\right),$$

where L is a function that returns the number of out-bound links of a node.

We specifically distinguish between *local* and *global* PageRank-score: local PageRank only uses connections/mentions of other users that have used the topic and made the topical "cut-off" described at the beginning of this chapter, while global PageRank uses connections/mentions from all users in the dataset.

We included this metric first and foremost because from the theory in Chapter 1, we see similarities in the random surfer model: when people are building

their interest network they most likely behave like a random surfer. When a user x is following a user a, who is following users $f(a) = \{y_1, y_2, \ldots, y_n\}$, it is very likely that there is a subset $I \subseteq f(a)$ that is interested in the same subject. When $y \in I$ is retweeted or mentioned by a, x might start following this user, or they could follow another random user that is also interested in the topic. In fact, TWITTER's follower recommendation system seems to at least take this "friends of your friends" information into account, although no official source could be found to support this. We add to our list:

8. Global PAGERANK: $p(x) = PR(x)$.
9. Local PAGERANK per topic: $p(x,t) = PR_t(x)$, which uses only the nodes/users that are present in the topic graph T_t.

4.2.4 HITS

Introduced in the same year as PAGERANK, Hypertext Induced Topic Selection (HITS) is an algorithm for internet importance that assumes a certain order in a graph introduced by Kleinberg [26]. It assumes there are two types of pages on the web: a hub is a node which consists mostly of a large collections of out-links, while an authority is a webpage that has little out-links, but many in-links from hubs.

We have included this algorithm because there are some similarities between the theories in Chapter 1 and HITS. If we assume that the hypothesis that two-step flow of information [23] is valid for social media networks, we must also assume there are "media" and "opinion leaders" that influence public opinion. Media has many links to the public and opinion leaders, who are in turn the authorities. If we hypothesize the definition of a hub to media, and the definition of authorities to opinion leaders, we get a pretty accurate picture of the way two-step flow could be at work in online social media. We again distinguish between the global and topical graph, and determine these metrics for both:

10. Topical Hub: the *hub* score of the node in the *topic graph*, denoted $h(x,t)$.
11. Global Hub: the *hub* score of the node in the *global graph*, denoted $h(x)$.
12. Topical Authority: the *authority* score of the node in the *topic graph*, denoted $a(x,t)$.
13. Global Authority: the *authority* score of the node in the *global graph*, denoted $a(x)$.

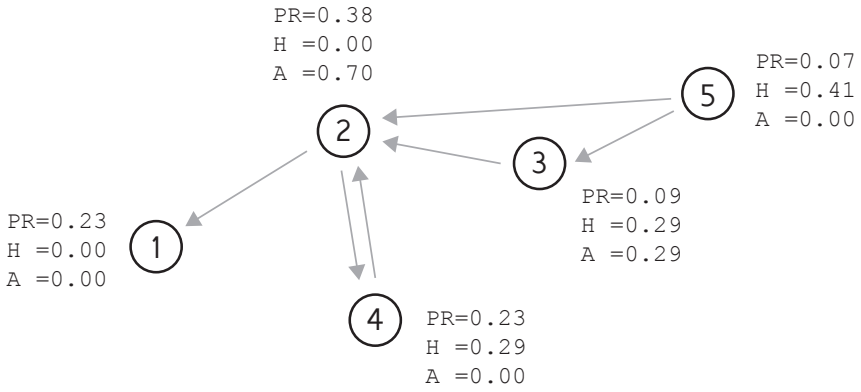

Figure 4.2: PageRank ($d = 0.15$) and HITS in our example.

4.2.5 Topic frequency

Topic frequency is a measure of a user's usage of the words specific to this topic. We use two metrics: one is the frequency of use of the topic in *all of the tweets of a user*, and the other is a weighted measure using the uniqueness of that word in the topic. The former metric is a simple addition of the number of tweets that are *on-topic*, where the latter uses an algorithm called tf-idf (term frequency-inverse document frequency) [22]. This is a measure that is used to weigh the author's use of a word to the uniqueness of the word in the entire collection. *Term frequency*, $\mathrm{tf}(t, d)$, is the number of times the term t has been used in a document d. Inverse document frequency is the inverse of the frequency of use of the term t in the entire collection of documents D and defined as

$$\mathrm{idf}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|},$$

with $|D|$ being the total number of documents in the dataset D and $|\{d \in D : t \in d\}|$ being the number of documents in which the term t is used. Then, the tf-idf is defined as

$$\mathrm{tf}{*}\mathrm{idf}(t, d, D) = \mathrm{tf}(t, d) * \mathrm{idf}(t, D).$$

The complexity of computation of these attributes is $O(n * p)$. So we add:

14. Frequency; or the number of messages from a user x on topic t, denoted by $f = |M_{tx}|$.

15. Tf-idf, as explained by the formula for $\mathrm{tf}{*}\mathrm{idf}(t, d, D)$, denoted by $\mathit{tfidf}(t, x)$.

4.2.6 Topical interest

To compensate for the inequality of the total number of messages sent by different users, we also use the ratio of *on-topic* messages and the total number of

messages sent by the users. The result is an attribute that gives the percentage of the tweets sent by this user that are related to a certain topic. Calculating this metric for every user and every message, the complexity is $O(n * p)$. We define:

16. Topical ratio: $r(t, x) = \dfrac{|M_{tx}|}{|M_x|}$.

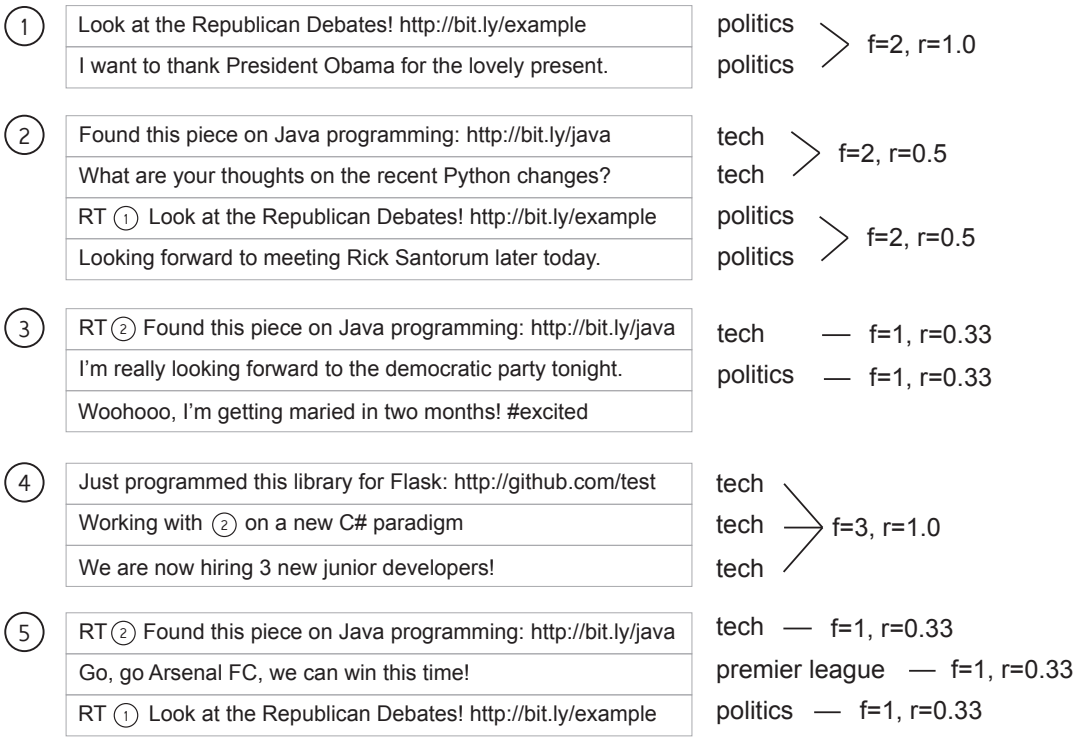

Figure 4.3: Frequency $f = |M_{tx}|$ and topical ratio r in our example.

4.2.7 Topic-sensitive PageRank

The topic-sensitive PageRank extends the original PageRank, described in Section 4.2.3, and adds a *personalization vector* based on the topical ratio, $r(t, x)$, from Section 4.2.6. Based on the work of Haveliwala [19], this variant of PageRank has another component: a vector of ratios indicating a user's usage of a topic. In our case, we will use the topical ratio $r(t, x)$, described above. This variant is explained by adjusting the equation for $PR(x)$ in Section 4.2.3 to the following equation of the topic-sensitive PageRank $TPR(x)$ for x, again with incoming links from nodes a, b and c:

$$TPR(t, x) = (1 - d)T(t, x) + d\left(\frac{TPR(a)}{L(a)} + \frac{TPR(b)}{L(b)} + \frac{TPR(c)}{L(c)}\right),$$

where $T(t,x)$ is the normalized vector of $r(t,x)$. We let:

17. Topical PageRank: $tpr(t,x) = TPR(t,x)$.

4.2.8 Topical following

For every user there are several measures that indicate the *topical interest* of their followers. That is, the sum of the topical interests of his or her followers, divided by the number of followers. This ratio indicates the level of interest in a certain topic of his or her followers. If the ratio is low, the followers have either no significant interests or highly diffused interest. If on the other hand the ratio is high, the followers have roughly the same interest; we can therefore suggest that the relationship with the followers is somehow related to this topic. These are quite resource-intensive attributes, given they need to be calculated over the followers/friends of all users, using content of all followers/friends, resulting in a complexity of $O(n * m * p)$. We define:

18. Average follower frequency: $\mathit{ff}(t,x) = \frac{1}{|I_x|} * \sum_{j \in I_x} |M_{tj}|$.

19. Average topical ratio of followers: $fr(t,x) = \frac{1}{|I_x|} * \sum_{j \in I_x} r(t,j)$.

20. Ratio of topical followers / total followers: $ti(t,x) = \frac{|I_x \cap T_t|}{|I_x|}$.

21. Ratio of topical friends / total friends: $to(t,x) = \frac{|O_x \cap T_t|}{|O_x|}$.

4.2.9 Neighborhood size

Neighborhood metrics stipulate the size of the neighborhood network of the user; not only its direct followers, but also the followers of his/her followers, the followers of the followers of his followers, etc. The set of users with distance r to user i using the *in-degree* (thus going "against" the direction of the edges) is denoted by $N_r(i)$. In this way, $N_1(i)$ is represents the same set as I_x. We mostly pay attention to $N_2(i)$, or *the followers of the people who follow user i*. We also separate the direct neighborhood from the indirect neighborhood by defining $N_d(i) = N_2(i) \cap N_1(i)$. This metric constitutes the average number of followers for each of the current node's followers. This is an indication of whether this person is on the outside, or on the inside of the graph. The complexity is $O(n * \frac{m}{n} * m)$. We define:

22. Indirect neighborhood: $N_2(x) = |I_x \bigcup_{i \in I_x} I_i|$.

23. Followers of followers: $N_d(x) = |N_2(x) \backslash N_1(x)|$.

24. Average followers of followers: $\mathit{fi}(x) = \dfrac{N_d(x)}{|I_x|}$.

Metric	Symbol	Complexity	Graph-based	Content-based
Followers	$\|I_x\|$	$O(n)$	$\surd$	–
Friends	$\|O_x\|$	$O(n)$	$\surd$	–
Mean global retweets	$\overline{rt}(x)$	$O(n*p)$	–	$\surd$
Mean topic retweets	$\overline{rt}(t,x)$	$O(n*p)$	–	$\surd$
Mean global mentions	$m(x)$	$O(n*p)$	–	$\surd$
Mean topic mentions	$m(t,x)$	$O(n*p)$	–	$\surd$
PageRank	$p(x)$	$O(n)$	$\surd$	–
Local PageRank	$p(t,x)$	$O(n*p)$	$\surd$	$\surd$
Topical Hub	$h(t,x)$	$O(n*p)$	$\surd$	$\surd$
Global Hub	$h(x)$	$O(n)$	$\surd$	–
Topical Authority	$a(t,x)$	$O(n*p)$	$\surd$	$\surd$
Global Authority	$a(x)$	$O(n)$	$\surd$	–
Topical frequency	$\|M_{tx}\|$	$O(n*p)$	–	$\surd$
tf-idf	$\mathit{tfidf}(t,x)$	$O(n*p)$	–	$\surd$
Topical ratio	$r(t,x)$	$O(n*p)$	–	$\surd$
Topic-sensitive PageRank	$tpr(t,x)$	$O(n*p)$	$\surd$	$\surd$
Follower frequency	$\mathit{ff}(t,x)$	$O(n*m*p)$	$\surd$	$\surd$
Follower ratio	$fr(t,x)$	$O(n*m*p)$	$\surd$	$\surd$
Ratio of topical followers	$ti(t,x)$	$O(n*m)$	$\surd$	$\surd$
Ratio of topical friends	$to(t,x)$	$O(n*m)$	$\surd$	$\surd$
Neighborhood	$N_2(x)$	$O(n*m)$	$\surd$	–
Followers of followers	$N_d(x)$	$O(n*m)$	$\surd$	–
Avg followers of followers	$\mathit{fi}(x)$	$O(n*m)$	$\surd$	–

Table 4.1: Overview of attributes.

4.3 Target attributes

Target attributes are attributes that we would like to predict. In our case, we would like to predict the attributes that contribute to Definition 2 and Definition 3. Specifically, for Definition 2 we would like to predict the probability that a message from a user generated a high number of clicks, and for Definition 3

we would like to know the the attributes of a user that generated a high number of retweets on his or her messages.

4.3.1 Click analytics

This metric is our ground truth for influence as defined by Definition 2. For all messages in the topic graph, we looked for the occurrence of hyperlinks that are wrapped with `t.co` and link to a `bit.ly` address (see Section 2.7). These messages are relatively rare, but give a solid measure of influence of the person sending the message. For all of these messages, we have retrieved the number of clicks from `bit.ly` that originate from the `t.co` location, and thus the original message. It should be noted that retweets do not alter the `t.co` link, so clicks include those who originate from retweets of the message.

When targeting this attribute, we only consider the people that have sent at least one message with such a link. We also reasoned that it is preferable to consider persons which have a consistent number of clicks, so we also measure standard deviation of click data, to determine stability.

1. Total clicks: $c(t,x) = \sum_{c \in C_{tx}} |clicks(c)|$, where $clicks(c)$ is the function that returns the number of clicks on link c.
2. Mean clicks per message: $\overline{c}(t,x) = \frac{1}{|C_{tx}|} * c(t,x)$.
3. Mean clicks per message, per 1000 followers: $cpm(t,x) = \overline{c}(t,x) / \frac{|I_x|}{1000}$.

4.3.2 Retweets and mentions

Earlier papers use retweets and mentions as an indication of influence. So for comparison, we will also use retweet and mention target attributes as defined in Section 4.2.2 to test influence based on Definition 3. In [8], mentions and retweets were found to have a high correlation; we will use this observation by considering a mention as just as important as a retweet. So we let:

1. Total topical mentions: $m(t,x) = \sum_{i \in T_t} |\{j \in M_{ti} : x \in mentions(j)\}|$, where $mentions(j)$ is a function that extracts the set of users mentioned in messages j.
2. Total global mentions: $m(x) = \sum_{i \in G} |\{j \in M_i : x \in mentions(j)\}|$, where $mentions(j)$ is a function that extracts the set of users mentioned in messages j.
3. Mean topical retweets per message: $\overline{rt}(t,x) = \frac{1}{|M_{tx}|} * \sum_{m \in M_{tx}} |R_m|$.

4. Mean global retweets per message: $\overline{rt}(x) = \frac{1}{|M_x|} * \sum_{m \in M_x} |R_m|$.

5. Mean topical retweets per message, per 1000 followers: $rpm(t, x) = \overline{rt}(t, x) / \frac{|I_x|}{1000}$.

4.3.3 Correlation between clicks and retweets

Before we try to extract significant attributes with clicks and retweets as target attributes, it would be interesting to find out whether there is a correlation between clicks, retweets and mentions. In Table 4.2, we see the Pearson correlation (see Section 4.4.1) between these attributes and find similar values as in the TWITTER research by Cha et al. [8]: mentions and retweets seem to be highly-correlated. Additionally, there is less correlation between clicks and retweets, and clicks and mentions, although there seems to be some indication that in certain topics, retweets and mentions might have a relation with clicks.

	Politics	Tech	Obama
Clicks vs RT	0.27	0.09	0.57
Clicks vs Mention	0.33	0.05	0.44
RT vs Mention	0.62	0.54	0.60

Table 4.2: Correlation between target attributes.

4.4 Extracting significant attributes

Before we start analyzing the relation between our target attributes and our source attributes in Section 4.5, we can benefit greatly from reducing the number of source attributes to only attributes that contribute to explaining the target variables, because this allows for easier interpretation of the eventual model and removes redundant attributes. This process is called *attribute filtering* and we will consider several types. To assist us in this process, we will use the WEKA [18] toolset. WEKA ("Waikato Environment for Knowledge Analysis") is a comprehensive set of data-mining tools, which allows for easy experimentation with a dataset like ours. We have used two algorithms to extract the most significant attributes from the total set of attributes: CfsSubsetEval and Principal Component Analysis (PCA).

4.4.1 Attribute correlations

To get more information on the exact relations between the attributes and targets, we will use the *Pearson product-moment correlation coefficient* (also called Pearson's r) on the attributes we have described. Pearson's r takes value in the

range $[-1, 1]$, where -1 indicates a perfect inverse correlation, $+1$ a perfect correlation, and 0 no correlation. We calculate the correlation between each of the 23 attributes from Section 4.2, resulting in a (symmetric) correlation matrix, which indicates the correlation strength and direction of all attributes. The tables for the topics "Obama", "Tech" and "Politics" can be seen in respectively Tables A.1, A.2 and A.3.

We noticed that overall the correlations are low; what to consider an acceptable correlation is dependent on the context, or the research and the field of science; in fields such as biology, chemistry, etc. correlations of 0.95 can be considered weak, while in the social sciences correlations higher than 0.5 are considered strong [11]. The main cause is imperfect measuring equipment (data noise) and the complexity of the experiments. Human behavior is very hard to predict and while our 23 attributes should provide a good indication, they are by no means a complete representation, and exact behavior could be influenced by many other factors, such as sentiment, country, time of day, outside weather, etc.

It should be noted that these are correlations on the entire topic datasets, including the users who have no click data. This means that the correlation of the clicks are very skewed towards 0. Therefore, we have also included the correlation of the clicks attributes for only users with click data available. From this we can see the strongest correlation with regards the target of average clicks is with average topical retweets. This indicates there is a strong relation between the number of topical retweets a message receives and the number of times a message is clicked. Also, all attributes relating to the number of followers are correlated: Topical PageRank, PageRank, Mentions, Followers and Audience all have strong correlation.

4.4.2 CfsSubsetEval

CfsSubsetEval [17] evaluates the value of a subset of attributes by considering the predictive ability of each attribute. It also tries to minimize the amount of redundancy between the attributes, thus giving a subset of attributes with high correlation with the target attribute, but low inter-correlation. This offers an insight into which attributes are important, yet does not tell anything about which feature is most important or the exact relation with target attribute. The results of this analysis can be seen in Table 4.3. Cfs in CfsSubsetEval is an acronym for Correlation based Feature Selector. Hence, the indicated merit is the measure of correlation between the composite of the attribute subset and the target variable.

We can see that for click data, the HITS hub score $h(x)$, number of followers $|I_x|$ and global retweets $\overline{rt}(x)$ and mentions $m(x)$ are the only popularity attributes. The other attributes are related to the relation of the user's *followers* relation to the topic, rather than the relation of the user him-/herself with the topic. A combination of these two types is represented in the average topical retweets $\overline{rt}(t, x)$, which represents popularity of topical messages, which are

Topic	Target	Merit	Selected attributes
Politics	$\overline{c}(t,x)$	0.745	$ti(t,x)$ $h(x)$, $fr(t,x)$, $\overline{rt}(t,x)$
Politics	$\overline{rt}(t,x)$	0.360	$p(x)$, $ti(t,x)$
Tech	$\overline{c}(t,x)$	0.458	$h(x)$, $fr(t,x)$, $\overline{rt}(t,x)$
Tech	$\overline{rt}(t,x)$	0.454	$a(t,x)$, $p(x)$, $ti(t,x)$, $\overline{rt}(x)$
Obama	$\overline{c}(t,x)$	0.671	$h(x)$, $\lvert I_x \rvert$, $\overline{rt}(x)$, $\overline{rt}(t,x)$, $m(x)$
Obama	$\overline{rt}(t,x)$	0.403	$a(x)$, $p(x)$
Premier League	$\overline{c}(t,x)$	0.466	$ti(t,x)$, $fr(t,x)$, $m(x)$, $\overline{rt}(t,x)$
Premier League	$\overline{rt}(t,x)$	0.537	$tpr(t,x)$, $h(x)$, $a(x)$, $p(x)$, $p(t,x)$, $m(x)$, $\overline{rt}(x)$

Table 4.3: Results of CfsSubsetEval on topics.

apparently well received by the followers of the user. When we target retweets instead, we again see the combination of topical and popularity attributes. In these cases, we mainly see PageRank $p(x)$, global retweets $\overline{rt}(x)$ and HITS authority $a(x)$ as the popularity attributes, and topical followers ratio $ti(t,x)$ as a measure of topical interest.

Taking into account the various correlations found in Section 4.4.1, we can now filter our original list of attributes into the most significant attributes to explain the variance in the dataset. We have taken into account the frequency of occurrence of the attributes and their correlation found earlier in Section 4.4.1.

The most significant attributes were found to be:

- Authority $a(x)$
- Hub $h(x)$
- Global PageRank $p(x)$
- Ratio of topical followers $ti(t,x)$
- Follower ratio $fr(t,x)$
- Average retweets $\overline{rt}(x)$
- Average mentions $m(x)$
- Average topical retweets $\overline{rt}(t,x)$

4.4.3 Principal Component Analysis

To expand on the attribute correlations in Section 4.4.1, we will also perform Principal Component Analysis (PCA) on the given attributes. PCA is a statistical method introduced by Pearson [39] which uses orthogonal transformations

to convert our attributes into a set of uncorrelated variables (*principal components*). The number of principal components is less than the number of original attributes, thus creating a linear combination of attributes that are correlated. Note that while the CfsSubsetEval method used in Section 4.4.2 is optimized using a target attribute, PCA does *not* have the notion of target attributes. It purely tries to minimize variance in the dataset by analyzing the source attributes and generate a set of new (composite) attributes.

We can use the principal components to explain the variance in the source attributes. In all topic graphs we found a common pattern: the principal component with the highest eigenvalue was always related to the popularity of the user. This includes the number of followers, mentions, retweets, PageRank, etc. For example, for the topic "Politics", this component is defined as follows:

$$0.313a(t,x) + 0.308a(x) + 0.305p(x) + 0.302|I_x| + 0.293N_2(x) + 0.287N_d(x) \\ +0.282p(t,x) + 0.261tpr(t,x) + 0.22rt(t,x) + 0.179m(t,x)\dots$$

This principal component will be labeled the *popularity* component.

When looking at the other significant components, we also consistently found a principal component which is related to the topic use of the user and even more so, the topic use of his/her followers. Common attributes are topical ratio, frequency, average follower topical ratio, topical PageRank and average topic follower. This indicates that at least a portion of the variance in the dataset can be explained by looking at the topic of the messages and the consistent use of this topic by the user and his/her followers.

Again, from the "Politics" dataset:

$$0.452\mathit{ff}(t,x) + 0.451fr(t,x) + 0.358ti(t,x) + 0.355\mathit{fi}(t,x) + 0.283r(t,x) \\ +0.262|M_{tx}| + 0.179tpr(t,x) + 0.179h(t,x)\dots$$

This principal component we will label as the *topical* component.

In this section we have learned that we can reduce the number of attributes by using correlation-based algorithms. Furthermore, by using PCA, we have discovered there are a range of topical attributes that are relevant to the variance in the dataset.

4.5 Explaining target attributes

The final step in the analysis is using the (significant) attributes that we retrieved in Section 4.4 to explain the target attribute variance. This is the main goal of this research: identify which attributes contribute to our definitions of topical influence. This type of problem is also called classification: Which attributes classify the value of the target attribute?

This idea is frequently used to analyze customer behavior. For example, we might have an attribute that indicates whether a customer has bought a certain

product, and a list of attributes that contains properties of the customer: age, income level, sex, postal area, education, etc. In this scenario, we would like to know what the attributes of our buyer most likely are, so we can recognize them. In the case of the current research, we want to identify the most likely attributes of the person with high influence. It might for example be classified by followers, PageRank and topical ratio.

The attributes we will use for this classification are determined by our results from Section 4.4. We have the following two scenarios:

- Using the attributes as indicated to be significant in the CfsSubsetEval in Section 4.4.2.
- Using the principal components as generated during PCA in Section 4.4.3.

These two sets of attributes are suitable because Langley and Sage stipulate that in the training of naive Bayes classifiers, no redundant attributes should be used in order to achieve maximal predictive performance [30]. The algorithms used to remove the redundancy between attributes should therefore allow us to generate better results.

There are many known algorithms with varying strengths and weaknesses for this particular problem set. Most commonly, there are *decision trees*, which give a predicted output based on the evaluation of a tree structure. This allows for dependency of attributes. The actual evaluation done at the node level, as well as the generation of the tree, is subject to many different algorithms. Modern algorithms such as Random Forests [7] even use many decision trees (hence the name forest) and take the mode of the trees. We however, have one additional demand for an algorithm: the resulting classifier must be easily *interpretable*. That is, it should be able to explain the relation between the attributes and the classifier clearly and easily.

4.5.1 Naive Bayes classifier

Another class of classifiers are known as naive Bayes classifiers. These models are based on Bayes' theorem [3] which stipulates how to interpret the probability of a certain target attribute based on the probabilities of one or more source attributes. More specifically, this algorithm uses a probability model that posits that the occurrence of the target attribute is a function of probabilities of the source attributes. It is a supervised machine learning algorithm, which means that there is a training before we can test the classifier. During training, a prediction of the attributes values is made. Usually, a normal (Gaussian) distribution is used to estimate these values. During testing, the source attributes of the unseen instance are used to calculate the probability of the different possible outcomes of the target attributes. The instance is then classified as the class that has the largest probability.

What makes this classifier *naive* is that it assumes independence of variables. This means that all attributes directly contribute to the probability of

the outcome, regardless of any dependence between variables. For example, the probability of an object being a car, given the number of wheels, weight, size, material, etc, is (naively) assumed to be determined by all of these variables uniformly, with no dependence between them. Even though in reality, weight might very well be dependent on material and size. The advantage is that the model is very simple, the outcome easily interpretable, and the algorithm can be quickly evaluated even on large datasets. Despite its simplicity, it has been shown to have good results relatively complex applications such as e-mail classification, intrusion detection, pattern recognition and document classification.

First, we tried to classify the average clicks ($\overline{c}(t,x)$) and average clicks per 1,000 followers ($cpm(t,x)$) by using a naive Bayes classifier. Our aim is to minimize the error made in classification, yet keep the solution simple. 10-fold cross validation is used to test the results of the classifiers. We will use Cohen's kappa κ [10] as our measure of error. This statistic takes into account the chance of random assignment to any class of the target attribute, and the measure indicates all possibilities between completely random assignment ($\kappa = 0$) or a completely accurate assignment ($\kappa = 1$). Because there is no consensus on what level of κ should be considered significant, we only use it to compare the different classifiers in the current research and will not assign a subjective significance to specific scores. Also, we will show the *confusion matrix* where the predictions of the classifier are shown against the actual class of the test set instances. This will be useful in determining whether a specific class of target attributes are difficult to predict.

First, we use binning to separate the target attribute into 4 possible classes (denoted by a,b,c,d, in ascending order), for target $\ln(1 + \overline{c}(t,x))$. The results on the filtered attributes from Section 4.4.2 on "Politics" resulted in a classifier with $\kappa = 0.4465$ and the following confusion matrix:

```
  a  b  c  d   <-- classified as
 71  9  2  0 |  a = '(-inf-2.114378]'
 28 21  6  0 |  b = '(2.114378-4.228757]'
  2 12 23  5 |  c = '(4.228757-6.343135]'
  0  0  3  4 |  d = '(6.343135-inf)'
```

Next, we try the same classification target attribute, but now using the two PCA attributes from Section 4.4.3. This results in a classifier with $\kappa = 0.238$ and confusion matrix:

```
 72  6  4  0 |  a = '(-inf-2.114378]'
 39  5 11  0 |  b = '(2.114378-4.228757]'
 11 10 19  2 |  c = '(4.228757-6.343135]'
  0  1  5  1 |  d = '(6.343135-inf)'
```

When testing attributes on "Tech": $\kappa = 0.401$

```
   a   b   c   d   <-- classified as
```

```
215  16   0   0 |   a = '(-inf-2.64126]'
 66  69  21   0 |   b = '(2.64126-5.28252]'
  3  30   9   6 |   c = '(5.28252-7.923781]'
  0   2   4   0 |   d = '(7.923781-inf)'
```

When testing PCA on "Tech": $\kappa = 0.2165$

```
  a   b   c   d   <-- classified as
193  24   4   0 |   a = '(-inf-2.64126]'
 83  31  31   3 |   b = '(2.64126-5.28252]'
 19  11  12   3 |   c = '(5.28252-7.923781]'
  1   2   3   0 |   d = '(7.923781-inf)'
```

When looking at the classifier results on "Politics" in Table 4.6, we find that there is a positive influence from attributes such as PageRank, HITS and global retweets, but an only slightly increasing influence of topical attributes such as topical ratio of followers. However, we see that average topic retweets $\overline{rt}(t, x)$, being a measure of both popularity as topicality, has a consistent upward moment towards the higher classes. We see a similar pattern when using PCA attributes in Table 4.4, where clicks increase with popularity, while topical has a constant, somewhat erratic behavior.

From looking at the results on "Tech" in Table 4.7, we see an even more interesting topical pattern: while popularity attributes such as PageRank and HITS are still increasingly important, the topical attributes in this dataset have a *negative* relation to the number of clicks. Also, average topic retweets $\overline{rt}(t, x)$ shows the same positive influence pattern as in "Politics". When using PCA on the same dataset, we can see from Table 4.5 that it follows the same pattern: topical attributes are more important in the lower two classes than in the higher two.

When testing the influence of Definition 3, or the influence of *retweets*, we found very similar patterns. As one can see from Table 4.8, the topical attributes follow a similar, erratic behavior with large standard deviations. The classifier has an accuracy of $\kappa = 0.3961$, and the confusion matrix was found to be:

```
  a   b   c   d   <-- classified as
884  67   0   3 |   a = '(-inf-2.092772]'
167 126  21   9 |   b = '(2.092772-4.185544]'
 32  55  26  19 |   c = '(4.185544-6.278316]'
  1  11   6   6 |   d = '(6.278316-inf)'
```

4.6 Conclusion

Concluding, it was very difficult to find a reliable, easy to understand metric, especially for the higher classes of clicks and retweets. Also we could not find a direct topical relation between clicks or retweets in any of the topic graphs.

We even found a (very weak) negative relation between most of the user's topical attributes. However, we can see that one of the most linearly increasing attributes is average topical retweets $\overline{rt}(t, x)$, an attribute that is related to both popularity as well as topicality.

PC	0	1	2	3
Popularity	−0.9923	0.1851	2.4967	6.5570
Topical	0.3335	0.7875	0.8005	−0.8308

Table 4.4: NBC on Politics PCA.

PC	0	1	2	3
Popularity	−0.1874	1.0652	1.9453	4.3041
Topical	1.6119	0.5065	−0.4157	−1.4424

Table 4.5: NBC on Tech PCA.

Attribute	0	1	2	3
$fr(t,x)$				
mean	0.0129	0.0179	0.0201	0.0071
std. dev.	0.0203	0.0251	0.0238	0.0040
$a(x)$				
mean	0.0001	0.0002	0.0008	0.0019
std. dev.	0.0001	0.0003	0.0006	0.0008
$h(x)$				
mean	0.0001	0.0001	0.0002	0.0003
std. dev.	0.0002	0.0002	0.0002	0.0003
$p(x)$				
mean	0.0001	0.0002	0.0005	0.0014
std. dev.	0.0001	0.0002	0.0004	0.0012
$ti(t,x)$				
mean	0.3489	0.3908	0.3829	0.2513
std. dev.	0.2056	0.2398	0.2215	0.0901
$\overline{rt}(x)$				
mean	0.0069	0.0229	0.0831	0.2753
std. dev.	0.0166	0.0359	0.0824	0.2826
$m(x)$				
mean	0.0147	0.0540	0.1421	0.3464
std. dev.	0.0267	0.0611	0.1230	0.3270
$\overline{rt}(t,x)$				
mean	0.3294	6.5952	35.6487	144.4324
std. dev.	1.0783	8.9119	36.0996	165.8723

Table 4.6: NBC on attributes from "Politics".

Attribute	0	1	2	3
$fr(t,x)$				
mean	0.0309	0.0185	0.0118	0.0075
std. dev.	0.0252	0.0133	0.0100	0.0047
$a(x)$				
mean	0.0000	0.0002	0.0004	0.0009
std. dev.	0.0001	0.0003	0.0004	0.0013
$h(x)$				
mean	0.0001	0.0001	0.0001	0.0001
std. dev.	0.0001	0.0002	0.0001	0.0002
$p(x)$				
mean	0.0001	0.0001	0.0003	0.0006
std. dev.	0.0001	0.0002	0.0002	0.0010
$ti(t,x)$				
mean	0.7166	0.6388	0.5218	0.4094
std. dev.	0.2345	0.1991	0.2232	0.2061
$\overline{rt}(x)$				
mean	0.0052	0.0295	0.0590	0.0583
std. dev.	0.0116	0.0488	0.1067	0.0828
$m(x)$				
mean	0.0354	0.1062	0.2665	0.1742
std. dev.	0.0653	0.2671	0.4972	0.2195
$\overline{rt}(t,x)$				
mean	0.3675	9.8398	41.1162	129.9801
std. dev.	0.8230	20.4164	85.0027	104.1415

Table 4.7: NBC on attributes from "Tech".

Attribute	0	1	2	3
$fr(t, x)$				
mean	0.0143	0.0102	0.0059	0.0040
std. dev.	0.0262	0.0148	0.0057	0.0027
$a(x)$				
mean	0.0001	0.0006	0.0013	0.0017
std. dev.	0.0002	0.0007	0.0012	0.0011
$h(x)$				
mean	0.0001	0.0002	0.0002	0.0002
std. dev.	0.0002	0.0002	0.0002	0.0002
$p(x)$				
mean	0.0001	0.0004	0.0009	0.0014
std. dev.	0.0002	0.0005	0.0010	0.0012
$ti(t, x)$				
mean	0.3510	0.2870	0.2419	0.1931
std. dev.	0.2411	0.1787	0.1342	0.1077
$\overline{rt}(x)$				
mean	0.0100	0.0618	0.1333	0.1937
std. dev.	0.0250	0.1343	0.1679	0.1373
$m(x)$				
mean	0.0466	0.1796	0.3654	0.2820
std. dev.	0.0829	0.2813	1.2376	0.1756

Table 4.8: NBC on attributes from "Politics", targeting topical retweets.

Chapter 5

Composing a metric

Because we had a demand of using only easily interpretable classifiers, we can now start using the results from Section 4.5.1 and the most important attributes from Section 4.4 to build composite attributes that explain a significant amount of the target attributes. Feeding this new composite attribute back into the analysis explained in the previous chapter, gives us a predictor for influence we can then use on each and every user to give an indication of their topical influence.

The results from PCA in Section 4.4.3 give us a first indication of the ratio and combination of the attributes selected in Section 4.4.2. We identify two major components in the PCA attributes: the *popularity* attributes and the *topical* attributes, which include the topical attributes of the direct neighborhood. We interpret these principal components into the following two composite attributes for given user x and topic t, in which all attributes are standardized:

POPULAR $0.4 * p(x) + 0.3 * a(x) + 0.3 * m(t, x)$

TOPICAL $0.3 * r(t, x) + 0.35 * ti(t, x) + 0.35 * fr(t, x)$

The factors of the attributes are estimated from the eigenvalues in the principal components, but in future work can be optimized by using machine learning algorithms.

5.1 Feedback loop

The first way we tested these attributes is by using the same approach as in Section 4.5.1: training a naive Bayes classifier with these two attributes instead of the original set of attributes. On the "Politics" topic, this resulted in a classifier with $\kappa = 0.3169$. This classifier, the properties of which are shown in Table 5.1, is thus slightly *more* accurate than the model we gathered from the original set of filtered attributes, while being significantly simpler to interpret. What is also interesting is that the same characteristics of the original set still holds: while the topical attributes show a somewhat constant distribution,

the popular attribute clearly shows an increasing value towards the higher click classifications. When plotting the parameters of the classifiers in Figure 5.1, this trend can clearly be seen, although standard deviations are generally high. The classifier for the "Tech" topic, which can be found in Table 5.3, also maintains the same properties, although the classifier in this case is weaker than the original, with $\kappa = 0.2019$. We did not include the model for the topic "Tech" on the target of *cpm* because it was very unreliable: with $\kappa = 0.0330$, results where only slightly more accurate than random assignment. However, the *cpm*-based classifier for "Politics" was more accurate with $\kappa = 0.1295$. The classifier (Table 5.2) shows the positive relation with topic, while the popular metric changes to nearly constant. This trend can be seen in Figure 5.2.

Attribute	0	1	2	3
TOPICAL				
mean	0.0636	0.3144	0.4168	0.0254
std. dev.	0.7157	1.0795	1.1530	0.7204
POPULAR				
mean	−0.3318	−0.1469	0.4865	2.0915
std. dev.	0.1324	0.3794	0.6886	2.3682

Table 5.1: NBC on composite attributes on "Politics" $\overline{c}(t, x)$.

Attribute	0	1	2	3
TOPICAL				
mean	−0.0691	−0.0921	0.2659	0.4494
std. dev.	0.4615	0.3748	0.9870	1.2149
POPULAR				
mean	−0.2999	−0.2770	0.1728	−0.0103
std. dev.	0.1470	0.2305	1.0340	0.5689

Table 5.2: NBC on composite attributes on "Politics" $cpm(t, x)$.

5.2 Ranking correlations

The second method we used to test these metrics is by using Kendall's rank correlation τ [24] to compare the rankings produced by our attributes, with those from PageRank, Topical PageRank and in-degree in Table 5.4. This is similar to one of the evaluation used in TwitterRank [47]. This shows the relation

Attribute	0	1	2	3
TOPICAL				
mean	0.6059	0.2029	−0.1240	−0.4018
std. dev.	0.9297	0.7831	0.5256	0.2764
POPULAR				
mean	−0.1705	0.1595	0.4971	1.5374
std. dev.	0.3171	0.6643	0.7103	3.0082

Table 5.3: NBC on composite attributes on "Tech" $\overline{c}(t, x)$.

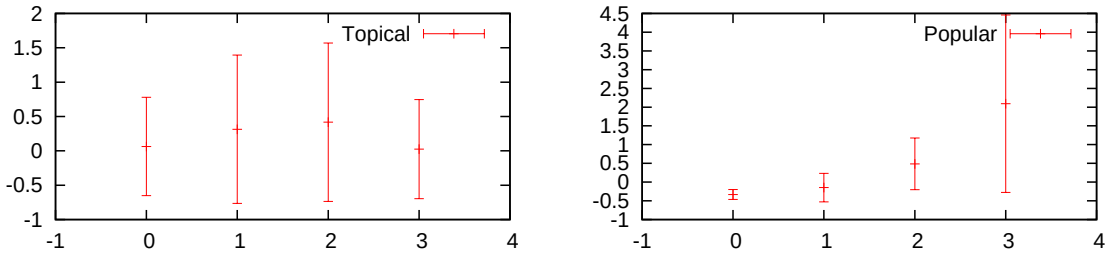

Figure 5.1: Topical (left) and Popular (right) distribution on "Politics". Average with standard deviation shown per class of $\overline{c}(t, x)$ (0-3).

between the attributes we have composed and some of the more traditionally used metrics. It can be seen that the rankings produced by POPULAR are largely correlated with the pure popularity measure of , whereas TOPICAL is much more aligned with topic-sensitive PageRank. In this way, the combination of POPULAR and TOPICAL have similar characteristics as TwitterRank.

5.3 Optimizing the metric

Finally, we decided to use an optimization algorithm to find an optimal composition of the attributes selected by CfsSubsetEval in Section 4.4.2 that maximizes the Kappa score. To this purpose, we have implemented a simple *genetic algorithm* [21], which is an approximation algorithm based on the theory of evolution. Genetic algorithms, in their simplest form, emulate the way nature uses evolution to incrementally improve a population of candidate solutions by random mutation and crossover through reproduction. They use a *fitness* function $f(x)$ as an indicator of the quality, and also the probability of survival of a candidate solution x (also called an "individual"). The major steps of the algorithm are:

1. Initialize population P_0 with n individuals

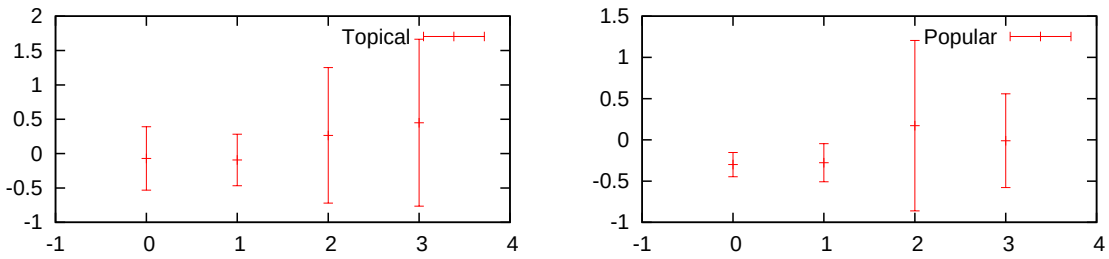

Figure 5.2: Topical (left) and Popular (right) distribution on "Politics". Average with standard deviation shown per class of $cpm(t, x)$ (0-3).

	Tech	Politics
PR vs In-Degree	0.7642	0.7554
PR vs TPR	0.5438	0.6147
POPULAR vs PR	0.8319	0.8610
POPULAR vs TPR	0.5291	0.6083
TOPICAL vs PR	−0.0800	0.0001
TOPICAL vs TPR	0.3154	0.3453

Table 5.4: Kendall τ of several metrics.

2. For generations $g = 0, 1, 2, \ldots, n$, evaluate population $\forall x \in P_g : f(x)$
3. Select two individuals a, b from P_g, preferring fit individuals
4. Reproduce a, b into individuals c, d, using mutation and/or recombination and add them to population P_{g+1}
5. Until some stop criterion s is met, repeat steps 2 through 4.

Our individuals consist of *two* composite attributes, composed of a total of n attributes, connected linearly with a weight for each of the attributes in the range $[0, 5]$ (e.g., $4 * a(x) + 3 * p(x)$ is an an attribute of an individual composed of attributes $a(x)$ with weight 4 and $p(x)$ with weight 3). As the fitness measure, we have continued using Cohen's kappa κ, the metric we have been using for the evaluation of all classifiers since Section 4.5.1. We have used both naive Bayes classifiers as well as C4.5 decision trees [40] as classifiers that produce the kappa metric, and selected the classifiers with the highest kappa.

We have limited our genetic algorithm to use only mutation. The first type of mutation occurs on the level of the individual by mutating the distribution of number of attributes in the composites, with $p = 0.05$ (e.g., an individual with one composite attribute of 3 attributes and one composite attribute of 4

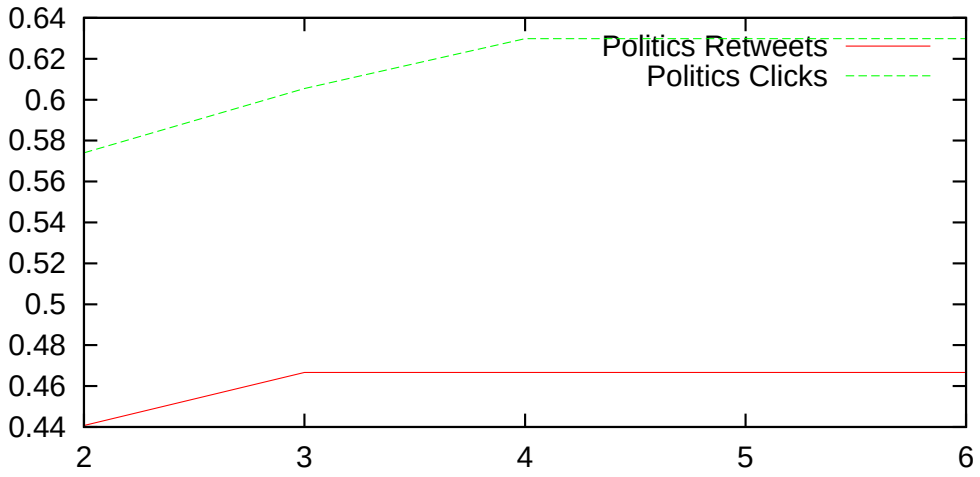

Figure 5.3: Kappa of best solution found by number of attributes in "Politics".

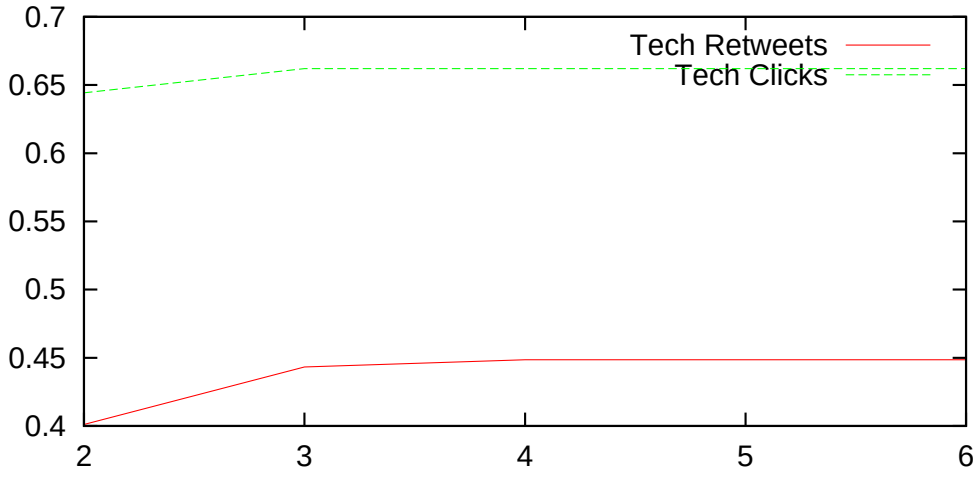

Figure 5.4: Kappa of best solution found by number of attributes in "Tech".

attributes might mutate to an individual with composites of respectively 1 and 6 attributes). The second, with probability $p = 0.1$, happens on the level of attributes which mutates both the weight and the selected attribute.

	Type	κ	Composite 1	Composite 2
Politics $\overline{rt}(t,x)$	C4.5	0.467	$3 * \overline{rt}(x)$	$4 * a(x) + 1 * m(x)$
Politics $c(t,x)$	C4.5	0.663	$4 * \overline{rt}(t,x)$	$4 * m(x) + 2 * h(x) + 1 * \overline{rt}(x)$
Tech $\overline{rt}(t,x)$	C4.5	0.469	$4 * p(x)$	$4 * \overline{rt}(x) + 1 * a(x)$
Tech $c(t,x)$	C4.5	0.661	$3 * \overline{rt}(t,x)$	$4 * m(x) + 4 * \overline{rt}(t,x)$
Tech $c(t,x)$ w/o $\overline{rt}(t,x)$	C4.5	0.444	$2 * \overline{rt}(x)$	$4 * ti(t,x) + 3 * m(x) + 2 * a(x)$
Politics $c(t,x)$ w/o $\overline{rt}(t,x)$	C4.5	0.458	$2 * m(x)$	$4 * a(x) + 3 * \overline{rt}(x) + 3 * h(x)$

Table 5.5: Composite metrics

To first find the optimal number of attributes that keep increasing the fitness of the best solution, we first ran the algorithm with several maximum number of attributes n, starting at $n = 2$. The results in Figure 5.3 and Figure 5.4 show that when targeting both average retweets $\overline{rt}(t,x)$ as average clicks $\overline{c}(t,x)$, the number of attributes that contribute to the best classifier are limited to four attributes in total.

The optimal composites as seen in Table 5.5 clearly show the influence of topical retweets. There are some popularity metrics such as $h(x)$ and $m(x)$. These popularity attributes are even more pronounced when we want to explain $\overline{rt}(t,x)$. Also, the models for $\overline{rt}(t,x)$ are far less accurate than the models for $c(t,x)$. In our observation, this is due to the large accuracy improvement the topical retweets provide. When removing $\overline{rt}(t,x)$, the accuracy of the classifier decreased to 0.444 — 0.458, as can be seen in the last two rows in Table 5.5, and consisted mostly of popularity attributes such as global retweets, HITS authority, global mentions. It seems the addition of topicality to the retweet interactions is accountable for most of the improvement beyond the popularity attributes.

Concluding, in this chapter, we have experimented with several compositions of the attributes that were found to be relevant in the previous chapter. We first found classifiers that relied mainly on popularity attributes. However, when using optimization techniques on the composite attributes, the average topical retweets was found to be the most important attribute, that improved the previous classifiers significantly. Together with some popularity attributes, it produced classifiers that were better than the classifiers in the original attribute space in Section 4.5.1.

Chapter 6

Conclusion

In this report, we have conducted a data-mining experiment on the social network Twitter, with the purpose of discovering the topical relations that tie users in the network. After collecting a subsample of the social graph using a Forest Fire algorithm and investigating it empirically, we first divided the graph into topical subgraphs by analyzing the use of certain predefined topics. The following data-mining process consisted of filtering and classifying the graph based on a large number of attributes of Twitter users. This showed us that the attributes based on topical content are much less important than the popularity when looking at the number of clicks they generated on posted links. However, the attribute of topical retweets was found to be predominant in all classifiers that were found. This attribute is a combination of popularity and topicality and seems correlated to the number of clicks in a certain topic. Note that this research is only meant to indicate correlation, and only suggests a relation between the attribute(s) and our specific definition of influence. The causation of the relation might have other reasons than the purely correlated attributes.

We believe that this is in accordance with and expands on earlier work. Bakshy et al. [1] found that the number of followers does not represent influence in the spreading of messages, and that large retweet cascades are originated mostly from many "less connected" ordinary users. Our finding that clicks correspond to high topical retweets, support this finding in that popularity is secondary to on-topic retweets and that the ability to generate topical activity is primary. Romero et al. [41] add that influence is determined by activity of followers, instead of passive attributes such as followers. This is confirmed by the fact that topical retweets are an activity from followers and not a passive metric such as followers or topical interest. Cha et al. [8] also suggest that followers are not the most important metrics, but content value is. Also they conclude that this influence is built over time. We believe topical retweets are an indication of content that fits well with the user's audience, which has been built over time, thus being a metric for both popularity, community and persistent content value.

In future work, we would like to use the methods used in this paper on a

larger scale to verify the results and investigate the patterns over time on the features we introduced (or other features that we have not considered). When verified, the results may be used in generating topical rankings based on the features that proved to be relevant. Among features we have not considered, time-based features (e.g., average follower growth, retweet average per month, etc.) would be an interesting addition to the current research. We could also see an influence ranking application (like Klout) use some of the methods to verify the significance of their influence signals.

Concluding, when looking for topical influence, we believe it is most helpful to primarily investigate the interactions the user causes on his topical messages, especially regarding retweets. Only then should popularity be taken into account.

Bibliography

[1] Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (2011), WSDM'11, ACM, pp. 65–74.

[2] Bastian, M., Heymann, S., and Jacomy, M. Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media* (2009).

[3] Bayes, T., and Price, R. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions Vol. 53* (Jan. 1763), pp. 370–418.

[4] Bharat, K., and Henzinger, M. R. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 1998), SIGIR'98, ACM, pp. 104–111.

[5] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research Vol. 3* (Mar. 2003), pp. 993–1022.

[6] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, No. 10 (Oct. 2008).

[7] Breiman, L. Random forests. *Machine Learning Vol. 45* (2001), pp. 5–32.

[8] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)* (May 2010), pp. 10–17.

[9] Clark, R. A., and Goldsmith, R. E. Market mavens: Psychological influences. *Psychology and Marketing Vol. 22*, No. 4 (2005), pp. 289–312.

[10] Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement Vol. 20*, No. 1 (Apr. 1960), pp. 37–46.

[11] Cohen, J. *Statistical power analysis for the behavioral sciences*, 2 ed. Lawrence Erlbaum, Jan. 1988.

[12] Dumais, S. T. Latent semantic analysis. *Annual Review of Information Science and Technology Vol. 38*, No. 1 (2004), pp. 188–230.

[13] Edelman, R. Edelman trust barometer. `http://www.edelman.com/trust/2009/`, 2009.

[14] Geman, S., and Geman, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 6 (Nov. 1984), pp. 721–741.

[15] Gladwell, M. *The Tipping Point: How Little Things Can Make a Big Difference*. Abacus, Feb. 2002.

[16] Godin, S. *Tribes: we need you to lead us*. Portfolio, 2008.

[17] Hall, M. Correlation-based Feature Selection for Machine Learning, 1998.

[18] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The WEKA Data Mining Software: An Update. *SIGKDD Explorations Vol. 11* (2009).

[19] Haveliwala, T. H. Topic-sensitive pagerank. *Proceedings of the Eleventh International World Wide Web Conference* (May 2003).

[20] Hill, S., Provost, F., and Volinsky, C. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science Vol. 21*, No. 2 (May 2006), pp. 256–276.

[21] Holland, J. H. *Adaptation in natural and artificial systems*. MIT Press, 1992.

[22] Jones, K. S. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation Vol. 28* (1972), pp. 11–21.

[23] Katz, E., and Lazarsfeld, P. *Personal Influence: The Part Played by People in the Flow of Mass Communications*. Free Press, 1955.

[24] Kendall, M. G. A New Measure of Rank Correlation. *Biometrika Vol. 30* (June 1938), pp. 81–93.

[25] King, R. S. The Top 10 Programming Languages. `http://spectrum.ieee.org/at-work/tech-careers/the-top-10-programming-languages`, Oct. 2011.

[26] Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *Journal of the ACM Vol. 46*, No. 5 (1999), pp. 604–632.

[27] Kleinberg, J. M. Challenges in mining social network data: processes, privacy, and paradoxes. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2007), KDD'07, ACM, pp. 4–5.

[28] Klout, Inc. Klout, March 2012. `http://www.klout.com`.

[29] Kwak, H., Lee, C., Park, H., and Moon, S. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web* (New York, NY, USA, 2010), WWW'10, ACM, pp. 591–600.

[30] Langley, P., and Sage, S. Induction of Selective Bayesian Classifiers. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence* (1994), Morgan Kaufmann, pp. 399–406.

[31] Langville, A. N., and Meyer, C. D. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.

[32] Leskovec, J., and Faloutsos, C. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2006), KDD'06, ACM, pp. 631–636.

[33] Leskovec, J., Kleinberg, J., and Faloutsos, C. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (2005), KDD'05, ACM, pp. 177–187.

[34] Madden, M. Privacy management on social media sites. Tech. rep., Pew Research Center, Feb. 2012. `http://pewinternet.org/Reports/2012/Privacy-management-on-social-media.aspx`.

[35] Meiners, N. H., Schwarting, U., and Seeberger, B. The Renaissance of Word-of-Mouth Marketing: A 'new' standard in twenty-first century marketing management?! *International Journal of Economic Sciences and Applied Research Vol. 3*, No. 2 (2010), pp. 79–97.

[36] Merriam-Webster. influence. `http://merriam-webster.com/dictionary/influence`, 2011.

[37] Milgram, S. The Small World Problem. *Psychology Today Vol. 2* (1967), pp. 60–67.

[38] Moon, J., and Moser, L. On cliques in graphs. *Israel Journal of Mathematics Vol. 3*, No. 1 (Mar. 1965), 23–28.

[39] Pearson, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Vol. 2* (1901), pp. 559–572.

[40] Quinlan, J. R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Series in Machine Learning. Morgan Kaufmann, 1992.

[41] Romero, D. M., Galuba, W., Asur, S., and Huberman, B. A. Influence and passivity in social media. In *Proceedings of the 20th International Conference Companion on World Wide Web* (2011), WWW'11, ACM, pp. 113–114.

[42] Scott, J., and Carrington, P. *The Sage Handbook of Social Network Analysis*. The Sage Handbook. SAGE, 2011.

[43] Smith, T., Coyle, J. R., Lightfoot, E., and Scott, A. Reconsidering models of influence: The relationship between consumer social networks and word-of-mouth effectiveness. *Journal of Advertising Research Vol. 47*, No. 4 (2007), pp. 387.

[44] The Nielsen Company. Trust in advertising. `http://nl.nielsen.com/site/documents/TrustinAdvertising_maart2009.pdf`, 2009.

[45] Twitter, Inc. REST API resources. `https://dev.twitter.com/docs/api`, 2012.

[46] Twitter, Inc. Streaming API methods. `https://dev.twitter.com/docs/streaming-api/methods`, 2012.

[47] Weng, J., Lim, E. P., Jiang, J., and He, Q. TwitterRank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining* (2010), WSDM'10, ACM, pp. 261–270.

[48] Wikimedia Foundation, Inc. Wikipedia: The free encyclopedia.

[49] Wu, S., Hofman, J. M., Mason, W. A., and Watts, D. J. Who Says What to Whom on Twitter. In *Proceedings of World Wide Web Conference)* (2011), WWW'11.

Appendix A

Attribute correlations

Table A.1: Correlations on topic "Obama". ($N = 654$)
Last two lines represents using only users with click data ($N_c = 76$).

	$r(t,x)$	$tpr(t,x)$	$fr(t,x)$	$\lvert M_{tx}\rvert$	$a(x)$	$h(x)$	$a(t,x)$	$h(t,x)$	$p(x)$	$p(t,x)$	$m(t,x)$	$ti(t,x)$	$\lvert I_x\rvert$	$\lvert O_x\rvert$	$to(t,x)$	$ff(t,x)$	$N_2(x)$	$N_d(x)$	$fi(x)$	$\overline{rt}(x)$	$m(x)$	$\overline{rt}(t,x)$	$rpm(t,x)$	$\overline{c}(t,x)$	$cpm(t,x)$
$r(t,x)$	1	0.36	0.05	0.86	0.05	0.03	0.09	0.08	0.12	0.12	0.29	0.13	0.07	0.09	0.08	0.04	0.01	0	0.04	0.07	0.02	-0.02	-0.02	0.01	0
$tpr(t,x)$	0.36	1	0.24	0.32	0.58	0.11	0.67	0.16	0.74	0.75	0.58	0.24	0.58	0.25	0.19	0.18	0.52	0.51	0.1	0.5	0.16	0.19	0.03	0.15	0.01
$fr(t,x)$	0.05	0.24	1	0.04	-0.14	0	-0.05	0.06	-0.09	-0.02	0.03	0.47	-0.14	-0.05	0.02	0.93	0.02	0.04	0.84	-0.08	-0.05	-0.07	-0.08	0	0.04
$\lvert M_{tx}\rvert$	0.86	0.32	0.04	1	0.11	0.05	0.14	0.08	0.18	0.17	0.32	0.06	0.13	0.16	0.1	0.02	0.06	0.05	-0.01	0.09	0.03	-0.01	-0.01	0	0.03
$a(x)$	0.05	0.58	-0.14	0.11	1	0.25	0.9	0.19	0.88	0.68	0.28	-0.18	0.94	0.28	0.24	-0.14	0.86	0.84	-0.16	0.63	0.31	0.4	0.12	0.21	0
$h(x)$	0.03	0.11	0	0.05	0.25	1	0.24	0.92	0.15	0.11	0.08	0.04	0.18	0.14	0.93	0.02	0.39	0.4	0.06	0.13	0.08	0.08	-0.03	0.06	0.01
$a(t,x)$	0.09	0.67	-0.05	0.14	0.9	0.24	1	0.28	0.82	0.77	0.33	-0.01	0.8	0.21	0.3	-0.08	0.81	0.8	-0.12	0.61	0.23	0.28	0.06	0.2	-0.01
$h(t,x)$	0.08	0.16	0.06	0.08	0.19	0.92	0.28	1	0.12	0.13	0.07	0.18	0.12	0.09	0.96	0.06	0.32	0.33	0.08	0.12	0.04	0.02	-0.06	0.04	0.02
$p(x)$	0.12	0.74	-0.09	0.18	0.88	0.15	0.82	0.12	1	0.85	0.38	-0.11	0.92	0.35	0.19	-0.11	0.74	0.7	-0.12	0.66	0.29	0.38	0.11	0.16	-0.01
$p(t,x)$	0.12	0.75	-0.02	0.17	0.68	0.11	0.77	0.13	0.85	1	0.38	0.04	0.66	0.25	0.18	-0.06	0.61	0.59	-0.08	0.54	0.18	0.23	0.04	0.15	-0.01
$m(t,x)$	0.29	0.58	0.03	0.32	0.28	0.08	0.33	0.07	0.38	0.38	1	0.04	0.31	0.15	0.1	0.02	0.22	0.21	-0.02	0.37	0.28	0.02	-0.01	0.15	0.03
$ti(t,x)$	0.13	0.24	0.47	0.06	-0.18	0.04	-0.01	0.18	-0.11	0.04	0.04	1	-0.21	-0.09	0.13	0.39	-0.12	-0.1	0.33	-0.12	-0.09	-0.1	-0.11	-0.03	0.06
$\lvert I_x\rvert$	0.07	0.58	-0.14	0.13	0.94	0.18	0.8	0.12	0.92	0.66	0.31	-0.21	1	0.41	0.2	-0.14	0.77	0.73	-0.16	0.65	0.4	0.38	0.11	0.18	-0.01
$\lvert O_x\rvert$	0.09	0.25	-0.05	0.16	0.28	0.14	0.21	0.09	0.35	0.25	0.15	-0.09	0.41	1	0.26	-0.05	0.24	0.21	-0.05	0.09	0.06	-0.01	-0.06	-0.02	-0.03
$to(t,x)$	0.08	0.19	0.02	0.1	0.24	0.93	0.3	0.96	0.19	0.18	0.1	0.13	0.2	0.26	1	0.02	0.37	0.37	0.03	0.15	0.06	0.02	-0.07	0.03	0.01
$ff(t,x)$	0.04	0.18	0.93	0.02	-0.14	0.02	-0.08	0.06	-0.11	-0.06	0.02	0.39	-0.14	-0.05	0.02	1	0.03	0.05	0.94	-0.09	-0.05	-0.06	-0.09	0	0.03
$N_2(x)$	0.01	0.52	0.02	0.06	0.86	0.39	0.81	0.32	0.74	0.61	0.22	-0.12	0.77	0.24	0.37	0.03	1	1	0.03	0.51	0.24	0.32	0.07	0.19	0
$N_d(x)$	0	0.51	0.04	0.05	0.84	0.4	0.8	0.33	0.7	0.59	0.21	-0.1	0.73	0.21	0.37	0.05	1	1	0.05	0.48	0.22	0.3	0.06	0.19	0
$fi(x)$	0.04	0.1	0.84	-0.01	-0.16	0.06	-0.12	0.08	-0.12	-0.08	-0.02	0.33	-0.16	-0.05	0.03	0.94	0.03	0.05	1	-0.11	-0.06	-0.07	-0.11	-0.01	0.01
$\overline{rt}(x)$	0.07	0.5	-0.08	0.09	0.63	0.13	0.61	0.12	0.66	0.54	0.37	-0.12	0.65	0.09	0.15	-0.09	0.51	0.48	-0.11	1	0.23	0.27	0.11	0.22	0.02
$m(x)$	0.02	0.16	-0.05	0.03	0.31	0.08	0.23	0.04	0.29	0.18	0.28	-0.09	0.4	0.06	0.06	-0.05	0.24	0.22	-0.06	0.23	1	0.1	0.03	0.04	0.02
$\overline{rt}(t,x)$	-0.02	0.19	-0.07	-0.01	0.4	0.08	0.28	0.02	0.38	0.23	0.02	-0.1	0.38	-0.01	0.02	-0.06	0.32	0.3	-0.07	0.27	0.1	1	0.57	0.03	0.01
$rpm(t,x)$	-0.02	0.03	-0.08	-0.01	0.12	-0.03	0.06	-0.06	0.11	0.04	-0.01	-0.11	0.11	-0.06	-0.07	-0.09	0.07	0.06	-0.11	0.11	0.03	0.57	1	0.01	0.06
$\overline{c}(t,x)$	0.01	0.15	0	0	0.21	0.06	0.2	0.04	0.16	0.15	0.15	-0.03	0.18	-0.02	0.03	0	0.19	0.19	-0.01	0.22	0.04	0.03	0.01	1	0.39
$cpm(t,x)$	0	0.01	0.04	0.03	0	0.01	-0.01	0.02	-0.01	-0.01	0.03	0.06	-0.01	-0.03	0.01	0.03	0	0	0.01	0.02	0.02	0.01	0.06	0.39	1
$\overline{c}(t,x)$	0.01	0.28	-0.01	0	0.64	0.17	0.5	0.1	0.5	0.42	0.23	-0.08	0.66	-0.09	0.1	-0.01	0.54	0.53	-0.09	0.57	0.44	0.56	0.13	1	0.39
$cpm(t,x)$	0.01	0.03	0.23	0.03	-0.02	0.03	-0.02	0.05	-0.04	-0.03	0.04	0.19	-0.03	-0.15	0.02	0.31	0	0	0.12	0.04	0.16	0.1	0.56	0.39	1

Table A.2: Correlations on topic "Tech". ($N = 4806$)
Last two lines represents using only users with click data ($N_c = 441$).

	$r(t,x)$	$tpr(t,x)$	$fr(t,x)$	$\lvert M_{tx}\rvert$	$a(x)$	$h(x)$	$a(t,x)$	$h(t,x)$	$p(x)$	$p(t,x)$	$m(t,x)$	$ti(t,x)$	$\lvert I_x\rvert$	$\lvert O_x\rvert$	$to(t,x)$	$ff(t,x)$	$N_2(x)$	$N_d(x)$	$fi(x)$	$\overline{rt}(x)$	$m(x)$	$\overline{rt}(t,x)$	$rpm(t,x)$	$\overline{c}(t,x)$	$cpm(t,x)$
$r(t,x)$	1	0.37	0.29	0.83	-0.04	-0.01	0	0.03	-0.03	-0.01	0.06	0.18	-0.03	0.04	0.06	0.27	-0.01	-0.01	0.04	-0.02	-0.02	-0.04	-0.04	-0.01	-0.01
$tpr(t,x)$	0.37	1	0.31	0.36	0.65	0.21	0.75	0.28	0.75	0.8	0.29	0.13	0.67	0.32	0.36	0.23	0.62	0.61	0	0.35	0.19	0.12	-0.04	0	-0.02
$fr(t,x)$	0.29	0.31	1	0.25	-0.07	0.01	-0.01	0.09	-0.07	-0.03	0.01	0.49	-0.08	0.03	0.09	0.92	0.02	0.03	0.26	-0.07	-0.08	-0.08	-0.07	-0.03	-0.03
$\lvert M_{tx}\rvert$	0.83	0.36	0.25	1	-0.01	-0.01	0.04	0.05	0	0.02	0.08	0.15	0.01	0.06	0.09	0.25	0.04	0.05	0.04	-0.01	-0.04	-0.03	-0.03	-0.02	-0.01
$a(x)$	-0.04	0.65	-0.07	-0.01	1	0.33	0.93	0.32	0.87	0.83	0.25	-0.11	0.94	0.23	0.32	-0.09	0.82	0.8	-0.07	0.46	0.32	0.27	-0.01	0.03	-0.01
$h(x)$	-0.01	0.21	0.01	-0.01	0.33	1	0.35	0.95	0.24	0.24	0.08	0.06	0.26	0.16	0.86	0.01	0.44	0.44	0.12	0.13	0.11	0.01	-0.05	0	-0.02
$a(t,x)$	0	0.75	-0.01	0.04	0.93	0.35	1	0.4	0.85	0.88	0.3	0.01	0.86	0.28	0.42	-0.02	0.85	0.84	-0.03	0.44	0.26	0.18	-0.03	0.02	-0.01
$h(t,x)$	0.03	0.28	0.09	0.05	0.32	0.95	0.4	1	0.26	0.29	0.11	0.17	0.26	0.22	0.92	0.09	0.47	0.48	0.16	0.12	0.08	-0.02	-0.07	-0.01	-0.03
$p(x)$	-0.03	0.75	-0.07	0	0.87	0.24	0.85	0.26	1	0.96	0.26	-0.08	0.92	0.28	0.3	-0.08	0.74	0.71	-0.06	0.49	0.3	0.26	0	0.03	0
$p(t,x)$	-0.01	0.8	-0.03	0.02	0.83	0.24	0.88	0.29	0.96	1	0.29	0	0.85	0.29	0.33	-0.04	0.75	0.73	-0.05	0.46	0.26	0.21	-0.01	0.02	-0.01
$m(t,x)$	0.06	0.29	0.01	0.08	0.25	0.08	0.3	0.11	0.26	0.29	1	0.01	0.27	0.04	0.1	0.01	0.2	0.19	-0.02	0.35	0.35	0.03	-0.01	0.01	0
$ti(t,x)$	0.18	0.13	0.49	0.15	-0.11	0.06	0.01	0.17	-0.08	0	0.01	1	-0.14	0	0.14	0.49	0.03	0.04	0.38	-0.11	-0.14	-0.14	-0.12	-0.03	-0.03
$\lvert I_x\rvert$	-0.03	0.67	-0.08	0.01	0.94	0.26	0.86	0.26	0.92	0.85	0.27	-0.14	1	0.32	0.32	-0.09	0.76	0.73	-0.1	0.51	0.37	0.32	0.01	0.03	0
$\lvert O_x\rvert$	0.04	0.32	0.03	0.06	0.23	0.16	0.28	0.22	0.28	0.29	0.04	0	0.32	1	0.45	0.04	0.25	0.24	-0.02	0.08	0.06	0.04	-0.02	0	-0.01
$to(t,x)$	0.06	0.36	0.09	0.09	0.32	0.86	0.42	0.92	0.3	0.33	0.1	0.14	0.32	0.45	1	0.1	0.47	0.47	0.09	0.12	0.08	-0.01	-0.06	-0.01	-0.03
$ff(t,x)$	0.27	0.23	0.92	0.25	-0.09	0.01	-0.02	0.09	-0.08	-0.04	0.01	0.49	-0.09	0.04	0.1	1	0.02	0.03	0.31	-0.08	-0.08	-0.08	-0.07	-0.03	-0.03
$N_2(x)$	-0.01	0.62	0.02	0.04	0.82	0.44	0.85	0.47	0.74	0.75	0.2	0.03	0.76	0.25	0.47	0.02	1	1	0.19	0.37	0.24	0.19	-0.04	0.02	-0.02
$N_d(x)$	-0.01	0.61	0.03	0.05	0.8	0.44	0.84	0.48	0.71	0.73	0.19	0.04	0.73	0.24	0.47	0.03	1	1	0.2	0.35	0.22	0.17	-0.04	0.01	-0.02
$fi(x)$	0.04	0	0.26	0.04	-0.07	0.12	-0.03	0.16	-0.06	-0.05	-0.02	0.38	-0.1	-0.02	0.09	0.31	0.19	0.2	1	-0.08	-0.09	-0.07	-0.07	-0.02	-0.02
$\overline{rt}(x)$	-0.02	0.35	-0.07	-0.01	0.46	0.13	0.44	0.12	0.49	0.46	0.35	-0.11	0.51	0.08	0.12	-0.08	0.37	0.35	-0.08	1	0.28	0.45	0.11	0.02	0
$m(x)$	-0.02	0.19	-0.08	-0.04	0.32	0.11	0.26	0.08	0.3	0.26	0.35	-0.14	0.37	0.06	0.08	-0.08	0.24	0.22	-0.09	0.28	1	0.12	0	0.01	0
$\overline{rt}(t,x)$	-0.04	0.12	-0.08	-0.03	0.27	0.01	0.18	-0.02	0.26	0.21	0.03	-0.14	0.32	0.04	-0.01	-0.08	0.19	0.17	-0.07	0.45	0.12	1	0.6	0.03	0.03
$rpm(t,x)$	-0.04	-0.04	-0.07	-0.03	-0.01	-0.05	-0.03	-0.07	0	-0.01	-0.01	-0.12	0.01	-0.02	-0.06	-0.07	-0.04	-0.04	-0.07	0.11	0	0.6	1	0.04	0.08
$\overline{c}(t,x)$	-0.01	0	-0.03	-0.02	0.03	0	0.02	-0.01	0.03	0.02	0.01	-0.03	0.03	0	-0.01	-0.03	0.02	0.01	-0.02	0.02	0.01	0.03	0.04	1	0.91
$cpm(t,x)$	-0.01	-0.02	-0.03	-0.01	-0.01	-0.02	-0.01	-0.03	0	-0.01	0	-0.03	0	-0.01	-0.03	-0.03	-0.02	-0.02	-0.02	0	0	0.03	0.08	0.91	1
$\overline{c}(t,x)$	-0.02	0.01	-0.09	-0.02	0.12	-0.01	0.06	-0.03	0.12	0.09	0.03	-0.1	0.14	-0.01	-0.03	-0.08	0.05	0.05	-0.07	0.09	0.05	0.46	0.3	1	0.91
$cpm(t,x)$	-0.01	-0.05	-0.09	-0.01	-0.03	-0.06	-0.05	-0.07	-0.02	-0.03	-0.01	-0.12	-0.02	-0.03	-0.07	-0.09	-0.06	-0.06	-0.08	0	-0.01	0.43	0.6	0.91	1

Table A.3: Correlations on topic "Politics". ($N = 1488$)
Last two lines represents using only users with click data ($N_c = 186$).

	$r(t,x)$	$tpr(t,x)$	$fr(t,x)$	$\lvert M_{tx}\rvert$	$a(x)$	$h(x)$	$a(t,x)$	$h(t,x)$	$p(x)$	$p(t,x)$	$m(t,x)$	$ti(t,x)$	$\lvert I_x\rvert$	$\lvert O_x\rvert$	$to(t,x)$	$ff(t,x)$	$N_2(x)$	$N_d(x)$	$fi(x)$	$\overline{rt}(x)$	$m(x)$	$\overline{rt}(t,x)$	$rpm(t,x)$	$\overline{c}(t,x)$	
$r(t,x)$	1	0.65	0.31	0.79	0.14	0.07	0.24	0.17	0.19	0.22	0.39	0.32	0.14	0.11	0.16	0.27	0.13	0.12	0.08	0.14	0.03	-0.02	-0.03	0	0.01
$tpr(t,x)$	0.65	1	0.33	0.67	0.62	0.15	0.72	0.21	0.73	0.74	0.53	0.23	0.61	0.2	0.22	0.29	0.55	0.54	0.04	0.48	0.18	0.19	0	0.06	0.04
$fr(t,x)$	0.31	0.33	1	0.29	-0.09	0.03	-0.02	0.1	-0.07	-0.01	0.04	0.53	-0.11	-0.04	0.05	0.93	0.02	0.03	0.64	-0.07	-0.06	-0.07	-0.08	-0.02	0.03
$\lvert M_{tx}\rvert$	0.79	0.67	0.29	1	0.2	0.1	0.32	0.2	0.25	0.29	0.48	0.26	0.2	0.17	0.19	0.24	0.2	0.19	0.04	0.18	0.04	-0.01	-0.03	-0.01	0.03
$a(x)$	0.14	0.62	-0.09	0.2	1	0.3	0.93	0.26	0.89	0.75	0.39	-0.14	0.95	0.23	0.29	-0.1	0.87	0.84	-0.1	0.59	0.33	0.32	0.03	0.12	0.02
$h(x)$	0.07	0.15	0.03	0.1	0.3	1	0.28	0.93	0.2	0.15	0.1	0	0.23	0.12	0.92	0.04	0.43	0.44	0.09	0.17	0.11	0.05	-0.05	0.06	0.02
$a(t,x)$	0.24	0.72	-0.02	0.32	0.93	0.28	1	0.3	0.86	0.83	0.45	-0.01	0.86	0.2	0.32	-0.03	0.85	0.83	-0.07	0.6	0.28	0.26	0	0.11	0.01
$h(t,x)$	0.17	0.21	0.1	0.2	0.26	0.93	0.3	1	0.19	0.18	0.13	0.11	0.19	0.11	0.96	0.11	0.4	0.4	0.13	0.16	0.08	0.01	-0.07	0.02	0.01
$p(x)$	0.19	0.73	-0.07	0.25	0.89	0.2	0.86	0.19	1	0.88	0.45	-0.08	0.93	0.27	0.23	-0.07	0.76	0.74	-0.08	0.62	0.31	0.36	0.05	0.09	0.01
$p(t,x)$	0.22	0.74	-0.01	0.29	0.75	0.15	0.83	0.18	0.88	1	0.41	0.03	0.75	0.23	0.22	-0.02	0.68	0.67	-0.05	0.55	0.23	0.28	0.02	0.08	-0.01
$m(t,x)$	0.39	0.53	0.04	0.48	0.39	0.1	0.45	0.13	0.45	0.41	1	0.03	0.45	0.25	0.16	0.03	0.29	0.28	-0.02	0.38	0.47	0.03	-0.02	0.06	0.02
$ti(t,x)$	0.32	0.23	0.53	0.26	-0.14	0	-0.01	0.11	-0.08	0.03	0.03	1	-0.18	-0.07	0.08	0.51	-0.08	-0.07	0.35	-0.11	-0.1	-0.11	-0.11	-0.04	0
$\lvert I_x\rvert$	0.14	0.61	-0.11	0.2	0.95	0.23	0.86	0.19	0.93	0.75	0.45	-0.18	1	0.33	0.25	-0.11	0.79	0.76	-0.12	0.63	0.4	0.34	0.05	0.11	0.02
$\lvert O_x\rvert$	0.11	0.2	-0.04	0.17	0.23	0.12	0.2	0.11	0.27	0.23	0.25	-0.07	0.33	1	0.28	-0.04	0.2	0.18	-0.04	0.07	0.06	-0.02	-0.05	-0.01	-0.03
$to(t,x)$	0.16	0.22	0.05	0.19	0.29	0.92	0.32	0.96	0.23	0.22	0.16	0.08	0.25	0.28	1	0.06	0.42	0.42	0.08	0.18	0.09	0.01	-0.08	0.01	-0.01
$ff(t,x)$	0.27	0.29	0.93	0.24	-0.1	0.04	-0.03	0.11	-0.07	-0.02	0.03	0.51	-0.11	-0.04	0.06	1	0.04	0.05	0.75	-0.07	-0.06	-0.07	-0.07	-0.02	0.02
$N_2(x)$	0.13	0.55	0.02	0.2	0.87	0.43	0.85	0.4	0.76	0.68	0.29	-0.08	0.79	0.2	0.42	0.04	1	1	0.1	0.49	0.26	0.29	0	0.13	0.03
$N_d(x)$	0.12	0.54	0.03	0.19	0.84	0.44	0.83	0.4	0.74	0.67	0.28	-0.07	0.76	0.18	0.42	0.05	1	1	0.12	0.47	0.25	0.28	0	0.13	0.03
$fi(x)$	0.08	0.04	0.64	0.04	-0.1	0.09	-0.07	0.13	-0.08	-0.05	-0.02	0.35	-0.12	-0.04	0.08	0.75	0.1	0.12	1	-0.09	-0.06	-0.06	-0.09	-0.02	-0.01
$\overline{rt}(x)$	0.14	0.48	-0.07	0.18	0.59	0.17	0.6	0.16	0.62	0.55	0.38	-0.11	0.63	0.07	0.18	-0.07	0.49	0.47	-0.09	1	0.27	0.32	0.12	0.09	0.01
$m(x)$	0.03	0.18	-0.06	0.04	0.33	0.11	0.28	0.08	0.31	0.23	0.47	-0.1	0.4	0.06	0.09	-0.06	0.26	0.25	-0.06	0.27	1	0.09	0.05	0.03	0.02
$\overline{rt}(t,x)$	-0.02	0.19	-0.07	-0.01	0.32	0.05	0.26	0.01	0.36	0.28	0.03	-0.11	0.34	-0.02	0.01	-0.07	0.29	0.28	-0.06	0.32	0.09	1	0.56	0.06	0.03
$rpm(t,x)$	-0.03	0	-0.08	-0.03	0.03	-0.05	0	-0.07	0.05	0.02	-0.02	-0.11	0.05	-0.05	-0.08	-0.07	0	0	-0.09	0.12	0.05	0.56	1	0.03	0.06
$\overline{c}(t,x)$	0	0.06	-0.02	-0.01	0.12	0.06	0.11	0.02	0.09	0.08	0.06	-0.04	0.11	-0.01	0.01	-0.02	0.13	0.13	-0.02	0.09	0.03	0.06	0.03	1	0.6
$cpm(t,x)$	0.01	0.04	0.03	0.03	0.02	0.02	0.01	0.01	0.01	-0.01	0.02	0	0.02	-0.03	-0.01	0.02	0.03	0.03	-0.01	0.01	0.02	0.03	0.06	0.6	1
$\overline{c}(t,x)$	-0.01	0.11	-0.06	-0.01	0.4	0.15	0.28	0.05	0.31	0.23	0.12	-0.12	0.43	-0.06	0.03	-0.06	0.38	0.37	-0.07	0.27	0.33	0.74	0.35	1	0.6
$cpm(t,x)$	0.02	0.07	0.09	0.04	0.07	0.05	0.01	0.02	0.02	-0.02	0.05	0	0.06	-0.13	-0.03	0.06	0.08	0.08	-0.03	0.04	0.2	0.38	0.69	0.6	1