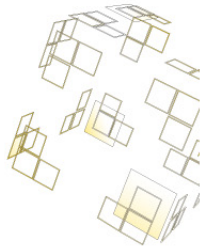


Thesis Applied Mathematics (Chair: Stochastic Operations Research)
Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)

Modelling trends in social media

Application and analysis using random graphs



Marijn ten Thij (s0135542)

Assessment Committee:

Prof. Dr. R.J. Boucherie (UT/SOR)
Dr. N. Litvak (UT/SOR)
T.M. Ouboter MSc. (TNO)
Dr. W. Kern (UT/DMMP)

August 23, 2013

UNIVERSITEIT TWENTE.

TNO

PREFACE

In front of you is my thesis, which was performed during his master Applied Mathematics at the University of Twente. This work was executed at the Performance of Networks and Systems (PoNS) group of TNO, the Dutch centre for applied research. I would like to use this moment to express my gratitude to several people who supported me throughout this project.

First, my colleagues at PoNS. I think I'm pretty lucky that I found such an open, welcoming and interested group to could perform my thesis project. This last quality is reflected by the large group of supervisors at PoNS. I want to thank Prof. Dr. Hans van den Berg and Dr. Daniël Worm for their guidance as secondary supervisors and above all I want to thank Tanneke Ouboter MSc. for her role as primary supervisor. Every time when I was to focused on a small part of the project, she helped me finding the right balance again through her fair and critical questions regarding my work. Another person that I want to thank specifically is ir. Dick van Smirren, thank you for all the Bila's and talks during my stay at TNO.

Second, I want to thank the staff of the Stochastic Operations Research chair at the University of Twente for welcoming me like a member of the staff instead of a student. But especially, thank you, Dr. Nelly Litvak for all the long conversations and your never ending optimism during these last months. It has been a great help in times when I wasn't convinced of my progress.

Finally I want to thank all the people, who read the concept version of this thesis and gave me feedback to build this final report. I hope that you will read this thesis with the same pleasure as I had constructing it.

Marijn ten Thij

CONTENTS

Abstract	1
1 Introduction	3
1.1 Related Work	4
2 Data	7
2.1 Tweet graphs	7
2.1.1 Retweet graph	7
2.1.2 Reply graph	8
2.1.3 Graph analysis	8
2.1.4 Centrality measures	9
2.2 Project-X dataset	9
2.2.1 Dataset analysis	10
2.3 Turkish-Kurdish dataset	12
2.3.1 Dataset analysis	13
2.4 WC speedskating dataset	14
2.4.1 Dataset analysis	14
3 Model	17
3.1 Random graph models	17
3.1.1 Preferential attachment model	17
3.1.2 Superstar model	18
3.2 Model definition	18
4 Model analysis	21
4.1 Growth of the graph	21
4.2 Component size distribution	27
4.2.1 The case $p=1$: a Pólya process	29
4.3 Size of the giant component	30
5 The model in practice	33
5.1 Deriving parameter estimators	33
5.2 Verification through simulation	33
5.3 Sensitivity analysis	34
5.4 Prediction using the estimates	36
5.4.1 Project-X dataset	36
5.4.2 Turkish-Kurdish dataset	39
5.4.3 WC speedskating dataset	40
6 Conclusions and discussion	43
6.1 Conclusions	43
6.2 Discussion	43
6.2.1 Improvements to the model	43
6.2.2 ‘Delayed’ superstar model	44
Bibliography	45

ABSTRACT

In this thesis we search for indicative properties to predict trending topics. We do this through an analysis of available tweet data, regarding an event that was discussed on *Twitter*. We analyse the progression of the tweets throughout time, using the retweet graph. We only take the structure of the graph into account for our analysis. We do not use any information with respect to the users that produced the tweets. We find that the average degree and the fraction of users that are part of the largest component are good indicators to consider for our purposes. We use this structural analysis to construct a random graph model, that captures the end progression of a retweet graph for a trending topic with only three parameters. We use this model to derive mathematical expressions for the indicators in our model. Moreover, we perform a sensitivity analysis of the model with respect to all three parameters. Also, we derive expressions to estimate the model parameters directly from available tweet data. Through these parameter estimates, we test our model on the analysed datasets. From these tests, we confirm the indicators and find an indication of trending behaviour. Thus, our model is well suited to simulate the progress of a retweet graph for a topic, in which a peak activity occurs.

CHAPTER 1

INTRODUCTION

In this thesis we analyse the progression of the retweet graph of three tweet datasets. Based on this analysis, we define a new random graph model, which is based on the superstar model by Bhamidi et al. [8]. We then analyse this model and obtain expressions, which we use to derive estimations for the parameters of the model. We also perform simulations using our model and compare these simulations to the actual data.

Currently, social media play an important role in our society. Everywhere we look, we are confronted with social media. Given the amount of interactions on these media, one can find large sets of data on how people communicate with each other on-line. One of the interesting topics in this setting is the detection of trends, since the topics people discuss online are an image of what interests the community.

During the last decade many studies have focused on social media. In this study, we focus on the microblogging platform *Twitter*¹. In *Twitter* users can post messages that consist of a maximum of 140 characters. These messages are called tweets. One can “follow” a user in *Twitter*, which places their messages in the message display, called the timeline. However, social ties are directed in *Twitter*, thus if user A follows user B, it does not have to mean that B follows A. People who “follow” you are called “friends”. We refer to the network of social ties in *Twitter* as the friend-follower network. Further, one can forward a tweet of a user to all their friends, which is called a retweet. Since *Twitter* is mainly used for the distribution of news [31], it is a natural platform for trends to emerge. *Twitter* uses the following definition for trending topics: “*Topics that are immediately popular, rather than topics that have been popular for a while or on a daily basis*”². In our study, we define a trending topic on *Twitter* to be a trend if the popularity of that topic is shaped by both on-line actions and real-life events, e.g. the calamity in Haren called “Project-X”. Many studies have focused on detecting these trends, for instance detecting emergencies [28], earthquakes [18,46,50], the flu [4,41,47], availability of on-line services [40] or events in sport events [32].

Early detection of trends in social media can be used for several purposes. Consider for instance the first mentioned application, detecting emergencies. In case of a disaster like a flood or an earthquake, warning people a few minutes in advance can save lives. Furthermore, looking at the spread of diseases, catching an epidemic early can produce a large gain. Moreover, in case of riots, law-enforcement can respond swiftly before matters spin out of control. Furthermore, the detection of trends also leads to an insight on the way trends are formed. Understanding this process gives a great advantage in several industries, e.g. in advertising.

In many current studies into trend behaviour, the focus is placed on content of the messages that are part of the trend, e.g. by Lehmann et al. in [36]. Our work focuses on the underlying networks describing the social ties between users of *Twitter*. In particular, we focus on the diffusion of topics through retweets. We aim to derive a mathematical model that can predict the probability of a trend occurring, using the current interactions in the social network. To the best of our knowledge, this is the first study that has chosen this approach to the trend detection problem. Therefore, the goal of our work is to answer the question:

How can we predict trending behaviour in Twitter using tweet based graphs?

To find an answer to this question, we investigate several characteristics of the tweet graphs, which we base on our analysis of the progression in the datasets. Using the characteristics we find in these datasets, we aim to derive a model that can simulate an outcome of the progression of a topic in *Twitter*. We try to answer the following questions:

¹www.twitter.com

²<https://support.twitter.com/articles/101125-about-trending-topics>

- i) Which graph characteristics can be used to model the progression of a topic?
- ii) Can we formulate a model, using these characteristics, to obtain a model that generates a similar final progression to that of a trending topic?
- iii) Can we use this model to estimate a probability that a topic becomes trending?
- iv) Can this estimate be validated by empirical data?

The lay-out of this thesis is as follows. First we describe the related research in Section 1.1. Second, we describe and analyse the datasets we used in Chapter 2. Then, we state our random graph model in Chapter 3 and analyse it in Chapter 4. Further, we use the theorems in Chapter 4 to derive predictions for the parameters of our model in Chapter 5. Finally, we formulate our conclusions and discuss further research opportunities in Chapter 6.

1.1 Related Work

The amount of literature regarding *Twitter* is vast. The overview we provide here is by no means complete. Many studies have been performed to determine basic properties of the so-called “Twitterverse”. Kwak et al. [31] analysed the follower distribution and found a non-power-law distribution with a short effective diameter and a low reciprocity. Furthermore they found that ranking by follower and PageRank both induce similar rankings. They also report that *Twitter* is mainly used for News (85% of the content). Huberman et al. [26] investigated interaction within the *Twitter* network. They found that the network of interaction within *Twitter* is not equal to the follower network, it is a lot smaller. Other aspects that are studied are trust by Adali et al. in [1] and credibility by Castillo et al. [14].

Another aspect of *Twitter* that is investigated is the notion of influence within the social network of *Twitter*. To rank influence, Yamaguchi et al. [54] define a ranking for *Twitter* users based on user-tweet graph analysis. Cano et al. [13] analyse social influence in the *Twitter* graph through a semantic profile. Ranking techniques are also used by Melville et al. [39] to find the viral potential of tweets. Furthermore, Hoang and Lim [23] define a mutual dependency model to define item and user virality and user susceptibility. They also derive an algorithm to determine the score related to these measures. Subbian and Melville [48] describe a predictive order-based rank aggregation to predict influence in a social network.

An important part of trending behaviour in social media is the way these trends progress through the network. To determine this, data needs to be extracted from available sources through a sample strategy. In [17], De Choudhury et al. search for the best sample strategy for the discovery of information diffusion in *Twitter*. Many studies have been performed on data that has been acquired from *Twitter*. For instance, Lerman and Ghosh [37] perform an empirical study on diffusion of information in *Twitter* and *Digg*. Bhattacharya and Ram [9] study the diffusion of news items in *Twitter* for several well-known news media and find that these cascades follow a star-like structure. This is confirmed by Cogan et al. in [16], where they investigate conversations on two subjects in *Twitter*. They find that the related tweet graphs structures range between stars and paths. Also, Zhou et al. [57] investigate the diffusion of information on *Twitter* using tweets on the Iranian election in 2009. They find that cascades tend to be wide, not too deep and follow a power law-distribution in their size. Furthermore, the cascade frequency does not depend solely on the amount of nodes in the cascade. In [25], Hsu et al. define several models for cascade size prediction. These models are based on participation of users, infection spread and community detection. In [27], Hui et al. define a model for the diffusion of actionable information, based on extracted cascades.

Another perspective on the diffusion of information in social media is obtained through analysing content of messages. Using this approach, Sadikov and Martinez [45] investigate the propagation of tags and URL's through *Twitter*. They found that tags tend to travel to more distant parts of the network and URL's travel shorter distances. Furthermore, Wu et al. [52] investigate the decay of content in *Twitter* and they find that short lived content is mostly negative, whereas longer lasting content is of a more positive nature. Romero et al. [44] analyse the spread mechanics of content through hashtag use. They derive probabilities that users adopt a hashtag.

One of the factors of a trend is its progression. Above we have seen several studies that modelled the progression as a whole. The following articles all focus on analysing or predicting one aspect of the progression of a topic, more precisely, the number and the behaviour of retweets. In [53], Xu and Yang analyse retweet behaviour to construct a model for retweet prediction. Hong et al. [24] add the content of the tweet as an extra factor for their prediction. Another model, formulated by Kuldeep et al. [29], uses a tweet topic model and the local retweet count of a user as factors for an apriori estimate of the number of retweets. Then, in [56], Zhang et al. describe several predicting models. The first model they describe uses the amount of followers that a user has to predict. The second model uses the separation between users. The third and last prediction method they define is based on a K -Nearest neighbour algorithm. Kupavskii et al. [30] design a machine learning algorithm to predict retweet cascade sizes.

Trends on *Twitter* can be divided in different categories. To achieve this, Zubiaga et al. [58] derive four different types of trends, using 15 features to make their distinction. The trends are distinguished as trends triggered by news, current events, memes or commemorative tweets. Lehmann et al. [36] study different patterns of hashtag trends in *Twitter*. They also observe four different classes of hashtag trends. A different approach is used by Budak et al. [12]. They extend the traditional definition of a trend into two subsets, defining trends to be *coordinated* in a connected community and *uncoordinated* if a user-graph of a trend consist out of multiple small graphs. They use their definition of trends to determine trend detection algorithms for both types of trends. Rattanaritnont et al. [43] find they can distinguish tweets based on four factors, which are cascade ratio, tweet ratio, time of tweet and patterns in topic-sensitive hashtags.

Detecting special topics in *Twitter* is a topic that has received much attention over the last years. There are several types of detection known in the literature. The first is First Story Detection (FSD), which aims to determine stories that are new in the datastream. Examples of this type of detection can be found in [2], where Allen et al. show that using Trend Detection and Tracking as a start to derive means for First Story Detection is impossible given the current tracking possibilities. Also Weng and Lee [51] define a system for event detection in the text stream of *Twitter* using clustering of wavelet-based signals. A second type of detection is peak or event detection. Below, we give some examples of this type of detection. Becker et al. [7] use an online clustering mechanism to find real-world events in a *Twitter* datastream. Ester et al. [19] derive several algorithms for trend detection in spatial databases. Popescu and Pennacchiotti [42] define several machine learning models to detect controversial events from *Twitter* streams. Lee and Kazutoshi [34] measure the regular tweet-activity in regions, which they then use as a benchmark to detect irregularities in these regions. Last, we mention anomaly detection. In [33], Lane and Brodley cast the anomaly-detection task in an instance-based learning framework. Lee and Xiang [35] use entropy as a measure for anomaly detection, in particular for intrusion detection. Sun et al. [49] define Neighbourhood formation in bipartite graphs which they then use to identify abnormal nodes in the graph.

Some authors combine trends and retweet analysis. For instance, in [22] Heard et al. define a two-step anomaly detection system. The first step is a Bayesian model to detect anomalous nodes. For the nodes that are detected, the second step is to induce the retweet subgraphs for these nodes. These subgraphs can then be analysed. Altshuler et al. [3] determine a lower bound for the success probability of an online-campaign, using a scale-free distributed follow network. In [55] Yang et al. use an adapted HITS algorithm and retweet graphs to detect trends. Li et al. [38] determine communities in a dynamically evolving environment using a probabilistic setting. This setting can also be used to determine the role of the nodes in the detected networks. Bhamidi et al. [8] propose a new random graph model based on the preferential attachment model, analysed in [6] by Barabási and Albert. They find that their model is well equipped to model giant components of retweet graphs.

CHAPTER 2

DATA

In the following chapter we describe the datasets we used in this study and analyse them. These datasets all contain tweets that have been acquired either using the *Twitter* Streaming API¹ or the *Twitter* REST API².

Using REST API one can obtain tweets or users from the *Twitter* database. For instance if you want the 20 most recent mentions of a specific user, you call the REST API using *GET statuses/mentions/timeline*.

The other option is using the Streaming API. This API filters tweets from the actual stream of tweets that *Twitter* parses during a day. Within this stream, one can filter on several objects. The first object is *follow*, in which a comma separated list of user ID's is specified. All tweets posted by the specified users are posted as output. The second object is *locations*, in which a geo-located square is specified. All tweets within this location are displayed as output. The third object is *track* for which a comma separated list of keywords is specified. All tweets that contain one of these keywords are posted as output.

There are several ways to use the API's from *Twitter*. We use *Python* and its module *tweetstream* in our study. The streaming API has a maximum amount of tweets that can be extracted from the total data stream of *Twitter*³, however we have not experienced any difficulties acquiring our tweets since the volume of tweets that we scraped stayed well below this threshold. We discuss several datasets, namely the *Project X*-dataset, the *Turkish-Kurdish*-dataset and the *WC speedskating*-dataset.

In our analysis of the datasets, we use some well known and some less known definitions. Therefore, we first cover the theoretic concepts that are considered required knowledge. Then, we describe the properties that we found in each dataset.

2.1 Tweet graphs

In *Twitter*, several types of graphs based on tweets can be constructed. All tweets that are contained in a dataset are depicted in some way in a so-called tweet-graph. This tweet-graph consists of two components, i.e. the retweet graph and the reply graph. Since we want to model the progression of a topic through the Twitter sphere, we are not interested in messages that are not retweeted or replied on. Therefore we can omit these messages from the graphs we analyse.

2.1.1 Retweet graph

We first define the retweet graph ($G_{\text{retweet}} = (V, E_{\text{retweet}})$). For this graph, we set V to be the set of users who either retweet a message or whose message is retweeted. Since there is not a specific format to formulate a retweet within *Twitter*, we restrict our analysis to a predetermined retweet setting. We detect two ways of retweet behaviour. One of them is starting a message with "RT: @A ...", which indicates that the user retweets a message by the user A. Another commonly used way of retweeting is ending a message with "... via @A". We search our dataset for these types of tweets and display an edge for every retweet in the dataset. We define an edge that indicates a retweet as follows:

$$\begin{aligned} e &= (u, v) \in E_{\text{retweet}} \Rightarrow \text{User } v \text{ retweets a tweet of } u. \\ V &= \{v \mid (u, v) \in E_{\text{retweet}} \cup (v, w) \in E_{\text{retweet}}\} \end{aligned}$$

¹<https://dev.twitter.com/docs/streaming-apis>

²<https://dev.twitter.com/docs/api/1.1>

³See <https://dev.twitter.com/docs/faq#6861>

We call the progression of a single message through the retweet graph a message tree. Notice that a user may interact in multiple message trees. See Figure 2.1 for an example.

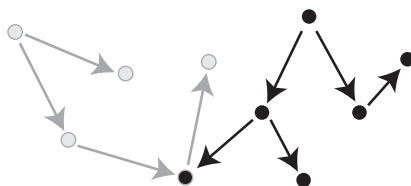


Figure 2.1: Example of a retweet graph.

2.1.2 Reply graph

We then consider the reply graph ($G_{\text{reply}} = (V, E_{\text{reply}})$). In this graph, we include all the conversations between users in our dataset. In *Twitter*, a user can reply to a tweet of another user. Let "Really looking forward to this weekend!" be a tweet of user A. Then user B can reply to this message by sending the tweet "@A me too!". We model this as an edge in the reply graph. The nodes in the reply graph depict the users that sent the tweets. Therefore we define V of our reply graph as the set of users that participate in one of the discussions. The directed edge set E contains all replies that are in our dataset. More formally:

$$e = (u, v) \in E_{\text{reply}} \Rightarrow \text{User } u \text{ replies to user } v.$$

$$V = \left\{ v \mid (u, v) \in E_{\text{reply}} \cup (v, w) \in E_{\text{reply}} \right\}$$

Since the reply graph of a topic is always much smaller than the retweet graph and replies mostly occur on a very local scale, we do not consider the reply graph in our analysis.

2.1.3 Graph analysis

There are many different ways to compare graphs with each other. In this section we explain which measures and distributions we use to analyse and compare the retweet graphs. Our definitions may vary from more well-known definitions, e.g. the definition of a giant component. This is a result of the fact that we use the program *Gephi* as a tool to analyse the dynamical progression of the graphs. Let \overline{G} denote the undirected version of the graph G . If both (u, v) and (v, u) are contained in G , they are combined to one edge (u, v) in \overline{G} . This means that both message trees depicted in Figure 2.1, indicated in grey and black, form one component. Before we introduce the different properties of the graphs that we consider, we first introduce two basic definitions.

Definition 2.1.1 (Giant component (GC)). The largest component of the graph \overline{G} is called the Giant component.

Note that, due to this definition, the component that is called the giant component can change over time. We shall see that this occurs in every dataset that we discuss in this thesis.

Definition 2.1.2 (Distance). Let $d(u, v)$ denote the distance between nodes u and v and let $P_{u,v}$ denote a path from u to v in \overline{G} , then

$$d(u, v) = \min_{P_{u,v} \in \overline{G}} \{|P_{u,v}| \mid P_{u,v} \text{ is a path from } u \text{ to } v \text{ in } \overline{G}\}.$$

One of the properties we derive for every graph is the degree distribution. Since we use directed graphs, we calculate the in-degree distribution, the out-degree distribution and the degree distribution. The second aspect is the distribution of component sizes. We derive the components of a graph using the *NetworkX* module of *Python*. To derive these size distributions, we consider the undirected version of the graph, \overline{G} .

2.1.4 Centrality measures

There are many ways to determine the importance of a node in a graph. The measures that are involved with these techniques are called centrality measures. In the following we describe several measures we use in our study.

Definition 2.1.3 (Eccentricity). The eccentricity, denoted by ϵ , of a node v in a graph $G = (V, E)$ is defined as the longest distance to another node in G ,

$$\epsilon(v) = \max_{u \in V} d(v, u).$$

Definition 2.1.4 (Graph diameter). In words, we define the diameter d of a graph G to be the length of the longest shortest path p in G . More formally

$$d = \max_{v \in V} \epsilon(v)$$

2.2 Project-X dataset

First, the *Project-X*-dataset contains tweets related to an event that occurred in the Netherlands. A *Facebook* event was made public, which eventually lead to large riots at the location of that particular event. The dataset was obtained by *Twitcident* for analysis of this phenomenon. A general overview of

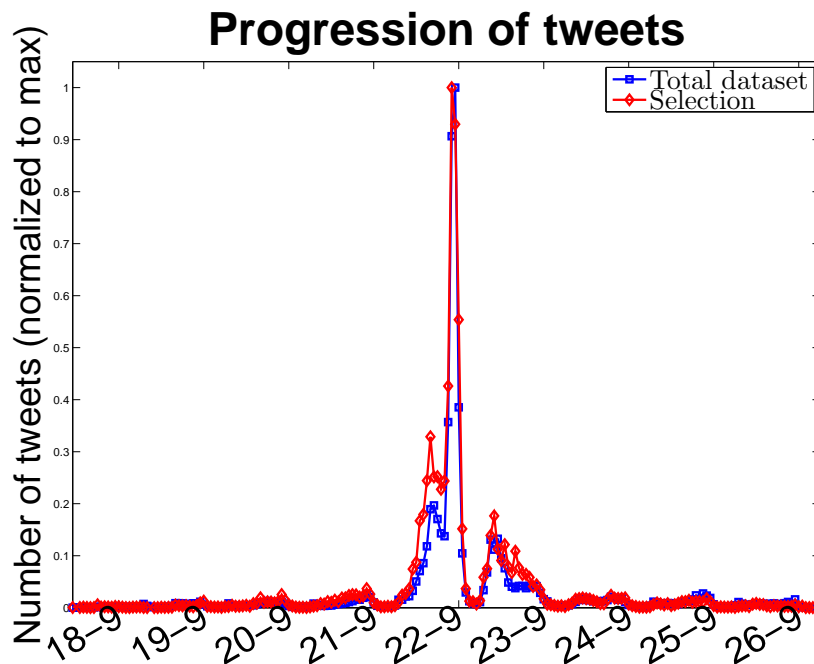


Figure 2.2: Normalized tweets for total en filtered dataset.

the dataset is given in the second column of Table 2.1. First the tweets with the keywords like "project x haren" and "feestbook feest" were acquired, using the earlier mentioned API's. At a later time, some extra keywords were added to obtain more tweets concerning the event. For more information on the retrieval of these tweets, we refer the reader to *Twitcident*⁴. Since this dataset contains many tweets, we decided to filter the dataset on two hashtags. These hashtags are *#projectx* and *#projectxharen*. The characteristics of this filtered dataset are shown in the last column of Table 2.1. Although this selection greatly reduces the number of tweets that we analyse, the progression of percentage of tweets per hour does not change significantly. It can therefore be concluded that the distribution of the volume of the

⁴www.twitcident.com

tweets in this dataset is similar for a narrow or a broad scope on the tweets of this subject. In Figure 2.2 we show the amounts of tweets per hour for both versions, normalized on the maximum number of hourly tweets.

	normal	filtered
first tweet	2012-09-17 11:36:18	2012-09-17 11:36:18
last tweet	2012-09-26 06:23:14	2012-09-26 04:32:15
tweets	739672	41283
users	255884	22072
retweets	368413	25832

Table 2.1: Characteristics of Project X dataset

2.2.1 Dataset analysis

In Section 2.2 we described the *Project-X* dataset. We use the information from this dataset as input for *Gephi*, a graph visualisation program. We analyse the growth of the retweet graph on an hourly scale. In Figure 2.3(a), we show the amount of tweets that appeared during an hour. The red line is the hourly amount of tweets, which corresponds to the right axis. On the left axis, we display the cumulative number of tweets with a blue line. In Figure 2.3(b) we depict the growth of the size of the node set and the edge set of the retweet graph. We see that both lines grow in a similar pattern throughout the progression.

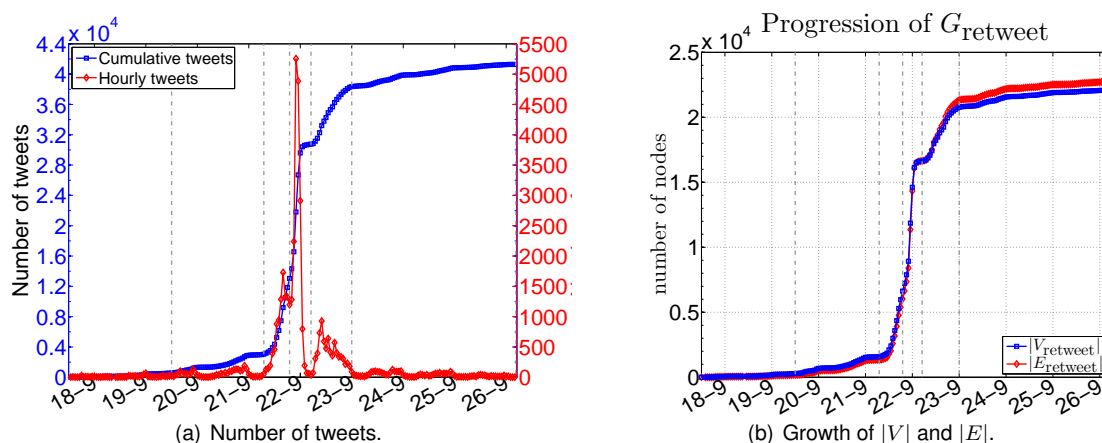


Figure 2.3: Progression of the retweet graph of *Project-X*.

The dash-dotted grey vertical lines in Figure 2.3 indicate times for which we analyse the properties of the graph more extensively. We have chosen six of such points in time. The first is an early progression of G (at 19-9-2012 12:00). Then, we analyse several moments during the peak activity. We chose the start of the peak (at 21-9-2012 7:00), just before the largest peak (at 21-9-2012 19:00), the end of the largest peak (at 22-9-2012 5:00) and the end of the peak activity (at 23-9-2012 0:00). Lastly we analyse G at the end of our dataset (26-9-2012 4:32). We chose these points in time because the graph has grown considerably in these intervals.

2.2.1.1 Growth of the giant component

We see that the giant component in the retweet graph contains the majority of vertices. This is in accordance to Ardon et al. [5], in which they define that a trend occurs when multiple communities merge in the progression of a subject. More intuitively, when components merge and become denser, the fraction between the number of edges and the number of nodes $\frac{|E|}{|V|}$ becomes larger. Therefore, we

may suspect that this fraction is an important indicator. We analysed the progression of this indicator, which is depicted in Figure 2.4(a). We see that for the giant component of the graph there is an interval when this fraction exceeds 1, indicated by the dash-dotted grey lines. Interestingly this is also the time when the fraction of nodes and edges that are contained in the giant component increases significantly, which can be seen in Figure 2.4(b). We shall call the moment that $\frac{|E_{GC}|}{|V_{GC}|}$ exceeds 1 the densification of the giant component.

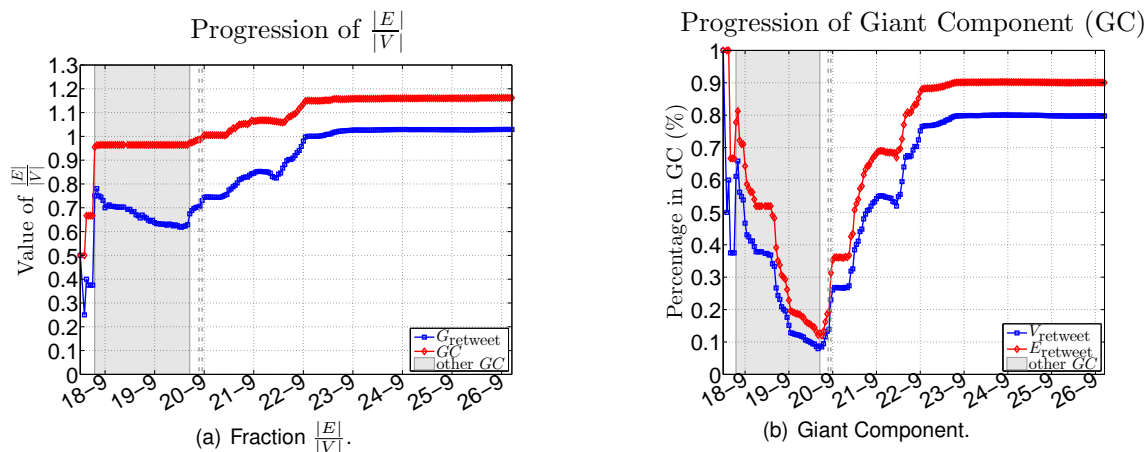


Figure 2.4: Progression for properties of the retweet graph of *Project-X*.

We then analyse the visualisations (see Figure 2.5) of this interval, which is between 19-9-2012 23:00 and 20-9-2012 0:00. The colouring of the nodes and edges is based on the components of the latest graph. In these figures, we see that several components, coloured in black, of the earlier graph (Figure 2.5(a)) are connected an hour later (in Figure 2.5(b)).

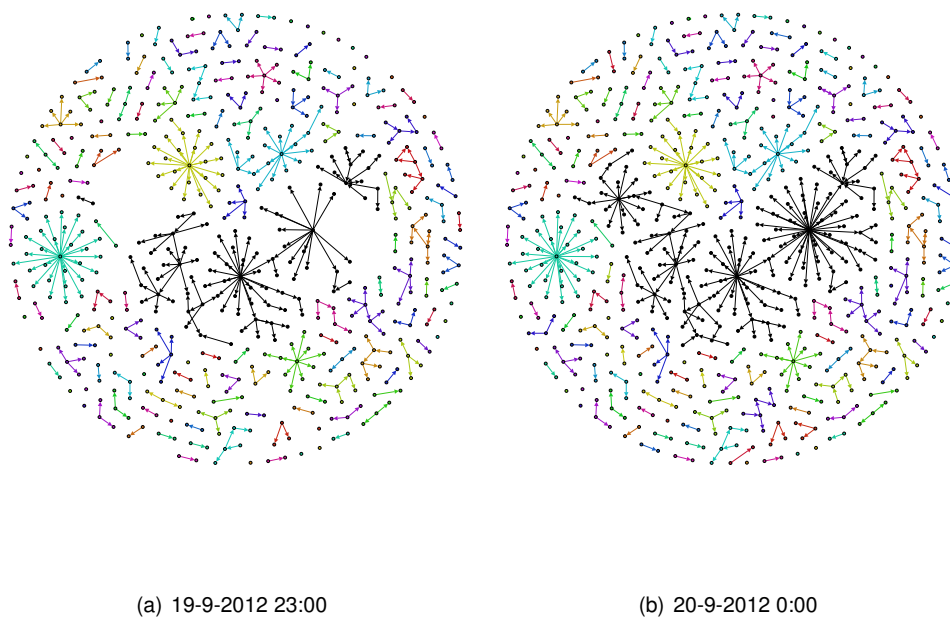


Figure 2.5: Densification of the retweet graph of *Project-X*.

Using *Gephi*, we can focus on multiple properties of the graph we analyse. We restrict our attention to the following properties and centrality measures: degree distributions (overall, in and out), component size distribution and eccentricity. Both Figures 2.5 and 2.6 indicate that the two largest components merge together forming a GC in the retweet graph. All other properties of the graph do not change significantly in this hour.

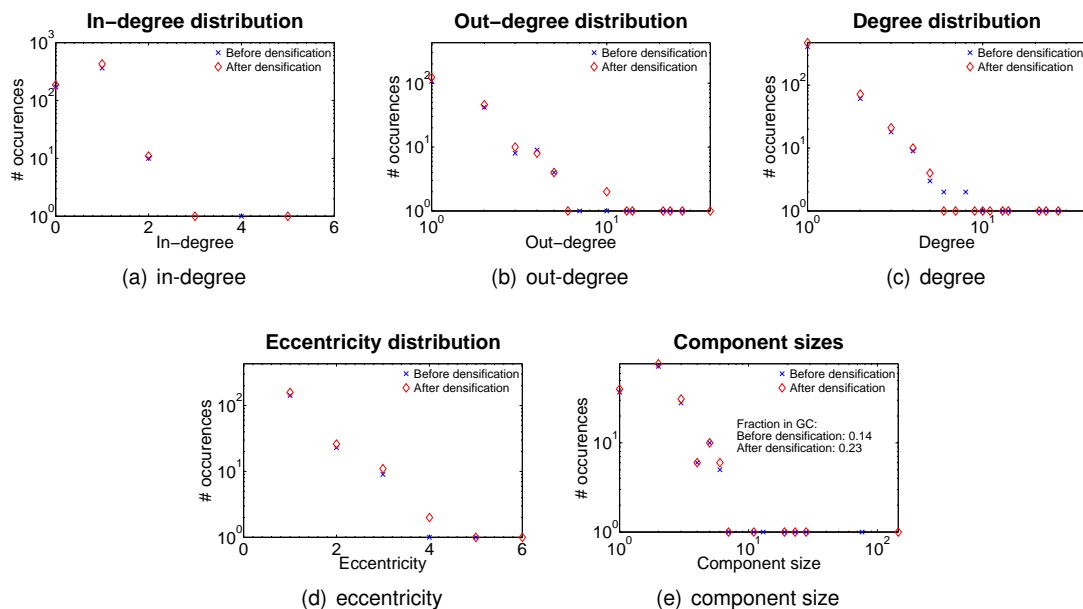


Figure 2.6: Distributions during the densification of the retweet graph of *Project-X*.

2.2.1.2 Growth during the peak

Then we switch to another change in the graph. We compare properties of the graph at the start of the peak with the properties at the end of the largest peak, see Figure 2.7. We see that for all degree distributions the slope is similar before and after the peak. This is easily explained by the large growth of the retweet graph in this time, which is directly evident when we look at Figures 2.3(a) and 2.3(b). However, there is one striking difference. The eccentricity score for the nodes after the peak are twice as high as the scores before the peak. Therefore, by joining the components, the paths that can be walked in the graph become twice as long.

Summarizing our findings, we see that our dataset has a densification of the amount of edges w.r.t. the amount of nodes in its progression. Also the eccentricity score, being the longest shortest path in the graph, increases after the peak. We thus want to derive a model that governs these aspects. With respect to the trend definition we use, this dataset can clearly be considered as a trend.

2.3 Turkish-Kurdish dataset

The second dataset contains tweets regarding demonstrations in the Netherlands that were related to the Turkish-Kurdish conflict in Turkey. To obtain the tweets from *Twitter*, these and other Dutch words were used: *koerden*, *turken*, *rellen*, *museumplein* and *amsterdam*. In Table 2.2 we describe the characteristics of the dataset. For a more detailed description of the dataset we refer the reader to [11] by Bouma et al.

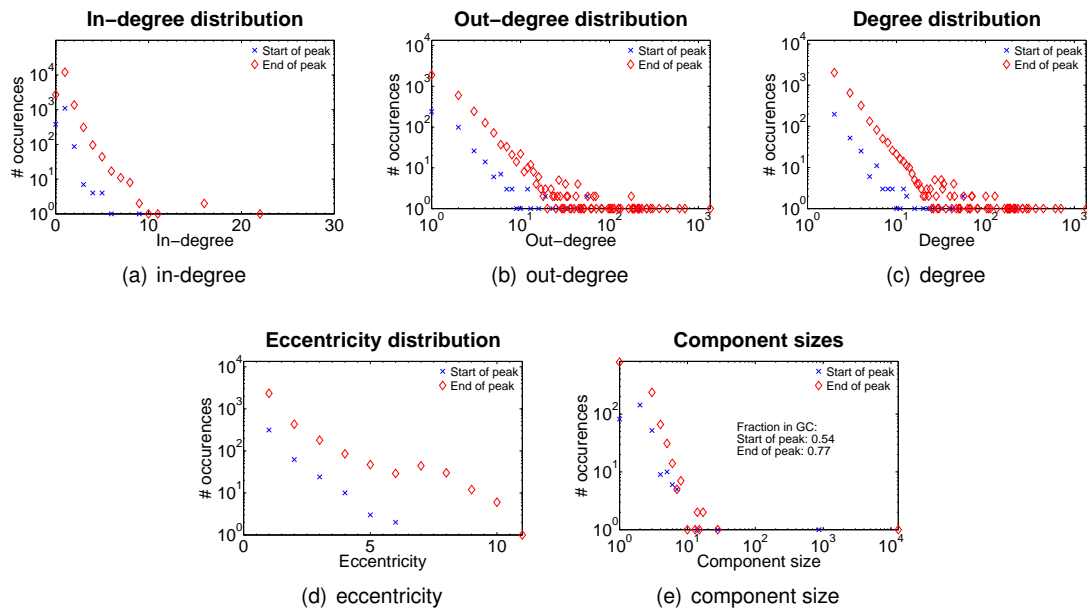


Figure 2.7: Properties during the peak.

	info
first tweet	2011-10-19 16:00:02
last tweet	2011-10-27 09:43:18
tweets	9655
users	3255
retweets	3547

Table 2.2: Characteristics of Turkish-Kurish dataset

2.3.1 Dataset analysis

If we compare Figures 2.8(a) and 2.3(a), we see that the *Turkish-Kurdish* dataset has many peaks, that mostly occur during the evening. Therefore, the development of the retweet graph is more regular than the development of the *Project-X* retweet graph.

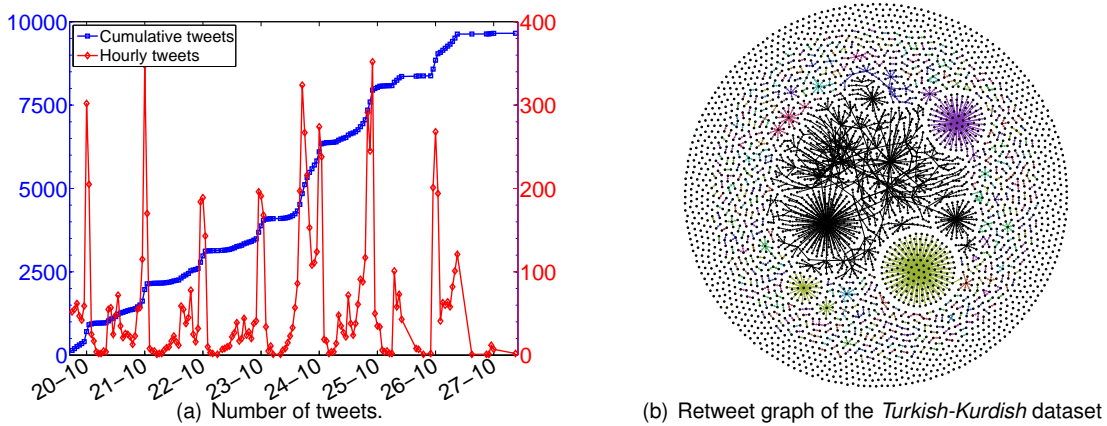


Figure 2.8: Progression of the retweet graph of the *Turkish-Kurdish* dataset.

Again, we analyse the growth of the giant component in this dataset. We find that the giant com-

ponent immediately has the densification property, mentioned earlier. However, on 21-10-2011 the progression of $\frac{|E|}{|V|}$ drops below 1 again, see Figure 2.9(a). This is a result of the fact that the largest component of the graph at that time is a different component than the eventual giant component. This component, located at the upper right of Figure 2.8(b), is indicated in purple. In Figure 2.9(b) we see the size of the giant component with respect to the size of the total retweet graph. Notice that these percentages drop both at the start of 24-10-2011 and at the start of 26-10-2011, indicating that at these times, the rest of the graph grows faster than the giant component. Furthermore, the giant component in this graph never contains more than 35% of the nodes in the graph, which is a clear distinction with the *Project-X* dataset. Given our trend definition, this dataset can be called a trend.

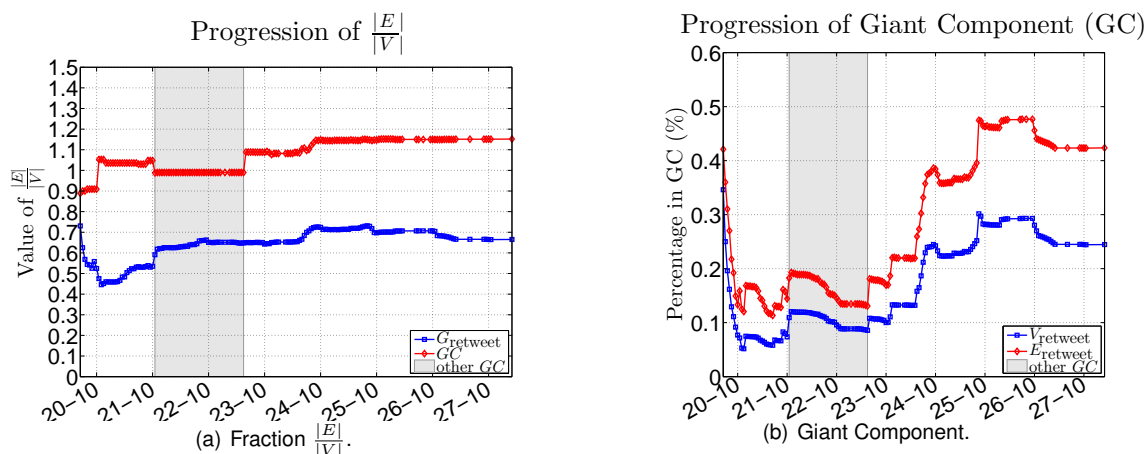


Figure 2.9: Progression for properties of the retweet graph of the *Turkish-Kurdish* dataset.

2.4 WC speedskating dataset

The last dataset consists of tweets that were scraped using the *Twitter-API* during the 2013 World Championship Speedskating for single distances, that took place in Sochi in Russia. We filter the *Twitter* stream on the hashtags *#wkafstanden*, *#sochi*, *#sotsji*. In Table 2.3 we show the characteristics of this dataset.

	info
first tweet	2013-03-21 08:54:59
last tweet	2013-03-25 08:46:50
tweets	3439
users	429
retweets	416

Table 2.3: Characteristics of WC speedskating dataset

2.4.1 Dataset analysis

The last dataset, containing the tweets regarding the WC speedskating 2013, is somewhat a combination of the first two datasets with respect to the distribution of the tweet volume. Although it has multiple peaks in its progression, as can be seen in Figure 2.10(a), one of these peaks is clearly the largest. Furthermore, the largest component, indicated in black in Figure 2.10(b), is very close to a tree-like structure.

This fact is also indicated in Figure 2.11(a), since the fraction $\frac{|E|}{|V|}$ just surpasses 1 at the end of the progression. This happens during the largest peak of the progression. Furthermore, at the beginning

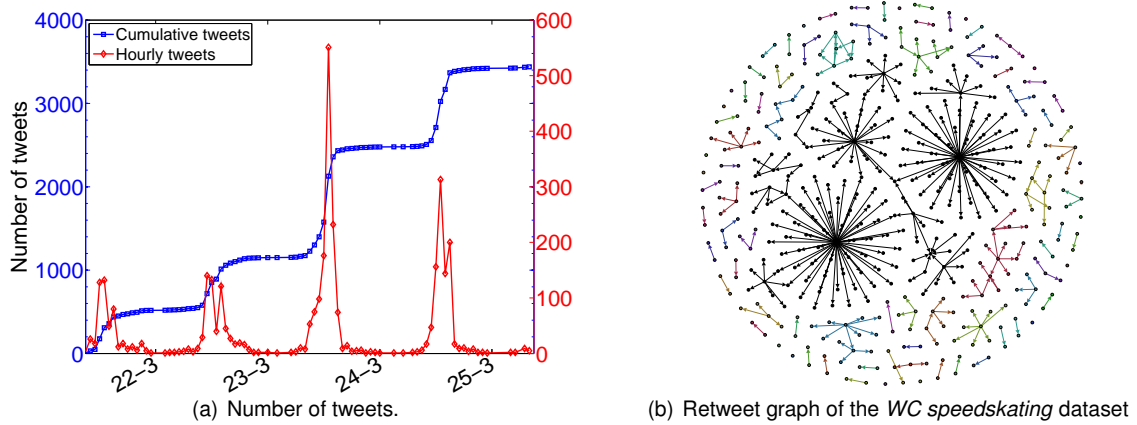


Figure 2.10: Progression of the retweet graph of the *WC speedskating* dataset.

of the progression the largest component is not the component indicated in black. It is the component indicated in light blue at the top of Figure 2.10. Lastly, the percentage of nodes that is contained in the giant component at the end of the progression is around 45%, which is similar to the *Turkish-Kurdish* dataset. Using our trend definition, we do not consider this dataset to contain a trend.

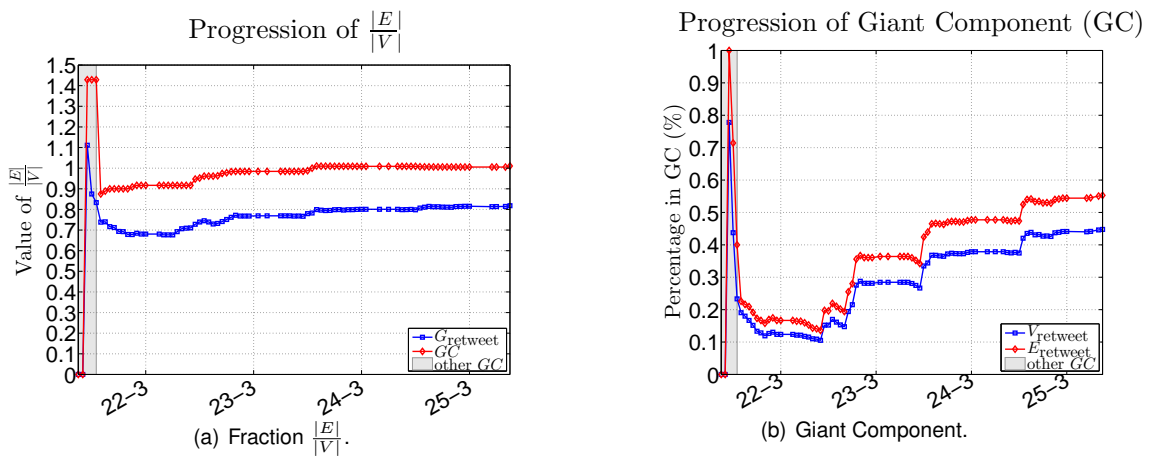


Figure 2.11: Progression for properties of the retweet graph of the *WC speedskating* dataset.

CHAPTER 3

MODEL

In this chapter we describe our model for trend analysis, which we base on our analysis of the *Project-X* dataset. We found that during the peak activity, the average degree of a node grows. Furthermore the fraction of edges per node is larger than one in the progression of the giant component of the retweet graph ($\frac{|E_{GC}|}{|V_{GC}|} > 1$). Recall that this observation is referred to as the densification of the retweet graph.

We use a dynamic random graph setting for our model, since it is most suitable for our situation. We base our model on the superstar model of Bhamidi et al. [8]. For the sake of simplicity of the model we neglect the friend-follower network of *Twitter*. Furthermore, every user can, in theory, retweet any message sent by a public user, which also supports our simplification. Before we state our model and derive its properties, we first introduce some background.

3.1 Random graph models

This section contains the models we use in our work and states the setup of these models and their properties.

3.1.1 Preferential attachment model

The first model we describe was extensively studied by Barabási and Albert [6] and it is referred to as the Preferential Attachment (PA) model. The model starts with an undirected graph G_0 of n_0 nodes. Then at every time step t a new node with m edges is added to the graph. The graph at time n in which we add m edges is denoted by G_m^n . Since we don't allow self-loops or multiple edges, it holds that $m \leq n_0$. The edges of the new node are connected to the existing graph using a PA scheme, that is the probability of attaching a new edge to a node i of degree δ_i is

$$\pi_i = \frac{\delta_i}{\sum_j \delta_j}.$$

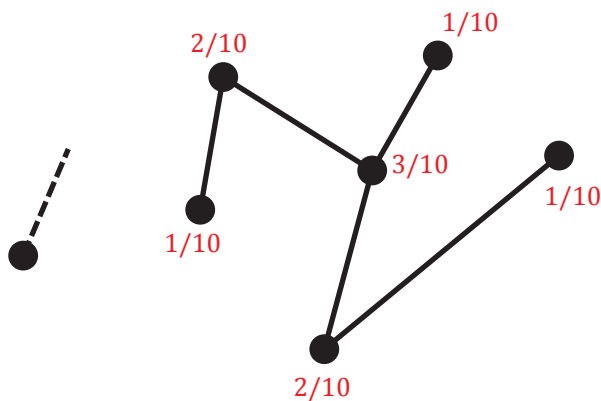


Figure 3.1: Example of preferal attachment scheme, where $m = 1$.

In Figure 3.1 we show an example of the different probabilities (indicated in red) with which the node with the dotted edge can join a specific node in the graph.

Recall that δ_v denotes the degree of a node v . The PA scheme, that starts with a single vertex with a loop, yields a graph with a power-law degree distribution, shown by Bollobás et al. in [10]. They derive this equation as follows. First the authors obtain the distribution of the sum of the first k degrees, which they use to obtain the expectation of $\#_m^n(k)$, defined as the number of vertices of the graph with in-degree k and total degree $d = k + m$. Finally they prove that $\mathbb{E}[\#_m^n(k) | G_m^t]$ is a martingale and use the Azuma-Hoeffding inequality to derive the following degree distribution:

$$P(\delta_v = d) = \frac{2m(m+1)}{d(d+1)(d+2)}.$$

If the network G_0 is connected, then the final graph will contain just one component. However, if G_0 consists of multiple components, there are several possibilities. If $m = 1$ the graph will remain disconnected. If $m \geq 2$ then the nodes that are added to the graph can connect the graph into one giant component. In our setting of retweet networks, it is most realistic to assume $m = 1$, since a person can only retweet one message at a time. However, in reality several components of the retweet graph can merge over time, thus this model needs to be adjusted to fit our needs.

3.1.2 Superstar model

Bhamidi et al. [8] define a model for undirected graphs that they call the superstar model. It seeks to model the retweet graph of a subject in the Twitter sphere. They found that this undirected graph can be modelled as a random graph using an extension of the PA model. At stage n , attach a new node v_n to the graph. We attach this node to the root v_0 , with probability q , or attach the node to a node from $\{v_1, \dots, v_{n-1}\}$ following the PA scheme, with probability $1 - q$.

3.2 Model definition

At the start of the progression, we have a graph G_0 . For now, we will assume that G_0 consists of a single node, however, this does not have to be the case. Since the graph evolves over time, we denote the graph process as $\{G_t, t = 0, 1, \dots\}$. At every time t there occurs one out of three arrival types to G_{t-1} , indicated in Figure 3.2, following the distribution given in Table 3.1.

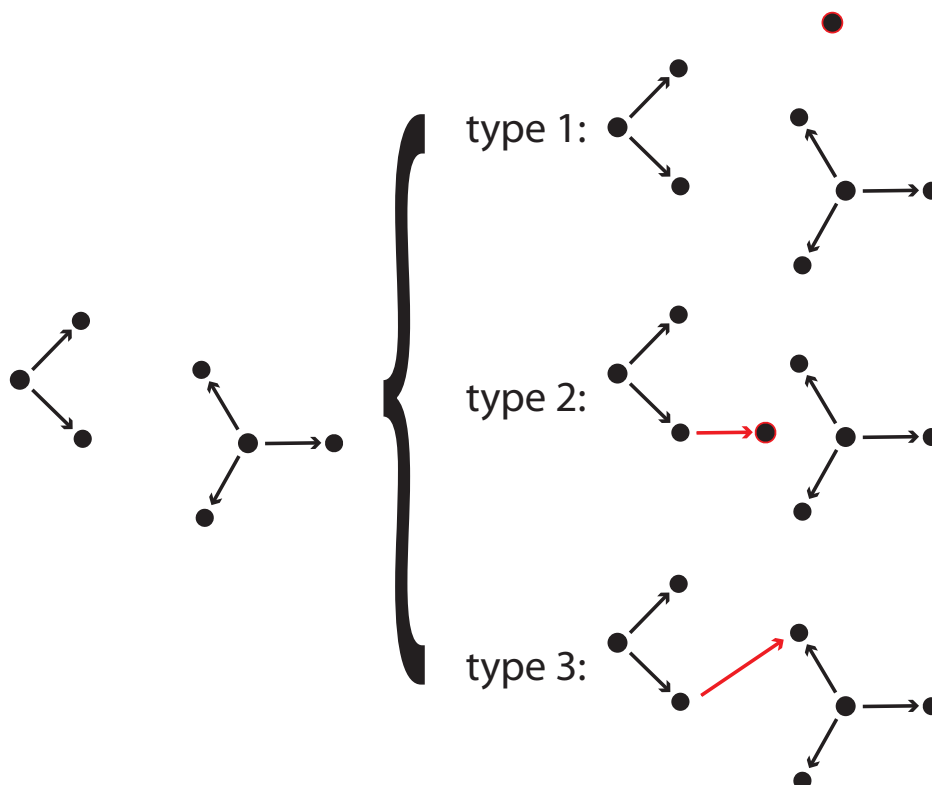


Figure 3.2: Arrival types to G_t .

For these arrival types, we use two parameters. The first parameter is λ , which is defined as the ratio of the rate of arriving messages (λ_m) to the rate of arriving retweets (λ_r). The second parameter is p , which is defined as the probability that a new retweet introduces a new node to the retweet graph.

Possible arrival types:	{	T1: a new message tree	w.p. $\frac{\lambda}{\lambda+1}$.
		T2: a retweet with a new user	w.p. $\frac{p}{\lambda+1}$.
		T3: a retweet with a known user	w.p. $\frac{1-p}{\lambda+1}$.

Table 3.1: Distribution of arrivals to G_t .

Once the type of a new arrival is known, we must determine the source node of a retweet (for both T2 and T3 arrivals) and choosing a target node of a retweet (for a T3 arrival).

First we discuss how a source node is chosen. Every message tree in G_t has a given probability that an arriving retweet joins its progression, this probability is proportional to the size of the message tree. Remember that these message trees do not have to be individual components in G_t , as is indicated in Figure 2.1. We define the probability that a retweet joins the message tree T_i (denoted by p_{T_i}) as

$$p_{T_i} = \frac{|T_i|}{\sum_{T_j \in G_t} |T_j|}$$

where we define $|T_i|$ as the number of nodes in the progression of that message tree. Therefore, if two message trees get joined during the growth of the graph, this fact does not influence their size. Notice that the probabilities p_{T_i} have to be updated each time an object (either a node or an edge) is added to the graph. After the message tree has been chosen, a node in this message tree is chosen as the source node following the superstar attachment scheme [8], using a third model parameter q . One could introduce a different superstar parameter q_{T_i} for every tree T_i . This way one could easily implement specific properties of the user that starts the message tree, e.g. his/her number of followers. For the sake of simplicity, we choose a uniform value of q for all message trees.

After a source node is chosen, a target node has to be chosen for a T3 arrival. We choose the target node by randomly selecting a target node from all nodes in G_{t-1} , with the exception of the earlier chosen source node, to prevent self-loops.

In Table 3.2 we present an overview of the declared parameters of the model.

notation	description
λ	rate of fraction between retweets and new messages
p	probability to join new edge to a new vertex
q	superstar parameter of the model

Table 3.2: Parameters of the model.

Parameters for extensions of the model	
notation	description
λ_r	Inter-arrival rate of new retweets
λ_m	Inter-arrival rate of new messages
q_{T_i}	extension to superstar parameter for message tree T_i

Table 3.3: Parameters for extensions of the model.

CHAPTER 4

MODEL ANALYSIS

In this chapter, we analyse the model stated in Chapter 3. We analyse several aspects of the model. First we derive some expressions that give an insight in the overall growth of the graph. Then we obtain an expression for the distribution of component sizes and derive expressions for the size of the giant component.

4.1 Growth of the graph

In this section we analyse how the graph grows over time. First, we derive expressions for the average degree in the graph G_t . We use two properties of our model that we formulate in two lemma's. The first of which provides an insight in the distribution of arrival types.

Lemma 4.1.1 (Distribution of arrival types). *Consider the graph G_t , then the distribution of the number of arrivals of type T_a , denoted by Y , is distributed as follows*

$$Y \sim \text{Bin}(t, \mathbb{P}(\text{arrival} = T_a)).$$

Proof. In the setup of our model an event happens at every time step. The probability that this event is of the type T_a is $\mathbb{P}(\text{arrival} = T_a)$. Since these events are i.i.d., it holds that $Y \sim \text{Bin}(t, \mathbb{P}(\text{arrival} = T_a))$. \square

In the next lemma, we analyse the the distribution of the number of T3 arrivals in between the arrivals of two nodes to the graph G_t .

Lemma 4.1.2 (Distribution in between new users). *We consider the time between the arrival of node $i - 1$ and node i to G_t . We state that the number of arrivals of T3 is distributed as the number of failures in,*

$$X_i \sim \text{Geo}\left(1 - \frac{1-p}{\lambda+1}\right).$$

Thus,

$$\mathbb{E}[\# \text{ T3 arrivals between } i \text{ and } i+1 \text{ nodes}] = \frac{1 - \left(1 - \frac{1-p}{\lambda+1}\right)}{1 - \frac{1-p}{\lambda+1}} = \frac{1-p}{\lambda+p}. \quad (4.1)$$

Proof. Consider that the $i - 1$ -th node just arrived at the graph. At every next time step we can add either a T1, T2 or T3 arrival to the graph. The probability of a T3 arrival occurring is $\frac{1-p}{\lambda+1}$ and therefore the probability that a T1 or T2 arrival occurs is $1 - \frac{1-p}{\lambda+1}$. Let us consider an arrival of T1 or T2 as a success. Then the number of T3 arrivals that occurs before a success is distributed as a geometric random variable with parameter $1 - \frac{1-p}{\lambda+1}$. The expectation of the number of failures of a geometric random variable with parameter \bar{p} is given by

$$\mathbb{E}[\# \text{ failures}] = \frac{1 - \bar{p}}{\bar{p}}.$$

Thus the number of T3 arrivals is distributed as a geometric random variable with parameter $1 - \frac{1-p}{\lambda+1}$ and has the expectation given in Equation 4.1. \square

The average degree is one of the aspects through which we give insight to the growth of the graph. In Theorems 4.1.3 and 4.1.4 we derive the expectation and the variance of the average degree in G_t , given the graph has just accumulated it's n -th node.

Theorem 4.1.3 (Expected average degree). *We consider a graph G_t , in which the n -th node has just arrived. Then*

$$\mathbb{E} \left[\frac{|E|}{|V|} \mid |V| = n \right] = \frac{1}{\lambda + p} - \frac{1-p}{n \cdot (\lambda + p)}. \quad (4.2)$$

Proof. We derive the expected average degree of a node v in G_t conditioned on the number of nodes in the graph.

$$\begin{aligned} \mathbb{E} \left[\frac{|E|}{|V|} \mid |V| = n \right] &= \frac{\mathbb{E}[|E| \mid |V| = n]}{n}, \\ &= \frac{\mathbb{E}[\# \text{ T2 arrivals on } [0, t] \mid |V| = n] + \mathbb{E}[\# \text{ T3 arrivals on } [0, t] \mid |V| = n]}{n} \end{aligned} \quad (4.3)$$

We investigate the time of arrival of the n -th node. Thus at that time, we had n arrivals of T1 or T2. Let X_j denote the type of the j -th arriving node. It follows that

$$X_j = \begin{cases} \text{T1 w.p. } \frac{\frac{\lambda}{\lambda+1}}{\frac{p}{\lambda+1} + \frac{\lambda}{\lambda+1}} = \frac{\lambda}{\lambda+p}. \\ \text{T2 w.p. } \frac{\frac{p}{\lambda+1}}{\frac{p}{\lambda+1} + \frac{\lambda}{\lambda+1}} = \frac{p}{\lambda+p}. \end{cases}$$

Therefore using a similar reasoning as Lemma 4.1.1, the distribution of the number of T1 arrivals follows a binomial distribution with parameters n and $\frac{\lambda}{\lambda+p}$ and the distribution of the number of T2 arrivals follows a binomial distribution with parameters n and $\frac{p}{\lambda+p}$. Consequently the expected number of T1 and T2 arrivals are

$$\mathbb{E}[\# \text{ T1 arrivals on } [0, t] \mid |V| = n] = n \cdot \frac{\lambda}{\lambda + p}. \quad (4.4)$$

$$\mathbb{E}[\# \text{ T2 arrivals on } [0, t] \mid |V| = n] = n \cdot \frac{p}{\lambda + p}. \quad (4.5)$$

Then we calculate $\mathbb{E}[\# \text{ T3 arrivals}]$. Using Lemma 4.1.2, the number of arrivals of T3 between i and $i + 1$ nodes, denoted by X_i , follows a geometric distribution with parameter $1 - \frac{1-p}{\lambda+1}$. Furthermore, we know that,

$$\mathbb{E}[\# \text{ T3 arrivals between } i \text{ and } i + 1 \text{ nodes}] = \frac{1-p}{\lambda + p}.$$

Since we have $n - 1$ of these transitions from 1 node to n nodes, we find

$$\mathbb{E}[\# \text{ T3 arrivals on } [0, t] \mid |V| = n] = (n - 1) \cdot \frac{1-p}{\lambda + p}. \quad (4.6)$$

Using (4.5), (4.6) and (4.3) we find

$$\begin{aligned} \mathbb{E} \left[\frac{|E|}{|V|} \mid |V| = n \right] &= \frac{n \cdot \frac{p}{\lambda+p} + (n - 1) \cdot \frac{1-p}{\lambda+p}}{n}, \\ &= \frac{1}{\lambda + p} - \frac{1-p}{n \cdot (\lambda + p)}. \end{aligned} \quad (4.7)$$

□

Theorem 4.1.4 (Variance of average degree). *We consider a graph G_t , in which the n -th node has just arrived. Then,*

$$\text{var} \left(\frac{|E|}{|V|} \mid |V| = n \right) = \frac{n \cdot p \cdot \lambda + (n - 1) (1 - p) (\lambda + 1)}{n^2 \cdot (\lambda + p)^2}. \quad (4.8)$$

Proof. Using Theorem 4.1.3, we find

$$\begin{aligned}
\text{var} \left(\frac{|E|}{|V|} \mid |V| = n \right) &= \frac{\text{var} (|E| \mid |V| = n)}{n^2}, \\
&= \frac{\text{var} (\#T2 + \#T3 \mid |V| = n)}{n^2}, \\
&= \frac{\text{var} (\#T2 \mid |V| = n) + \text{var} (\#T3 \mid |V| = n)}{n^2}. \tag{4.9}
\end{aligned}$$

We then define $\{\#T2 \mid |V| = n\} = Y$ and $\{\#T3 \mid |V| = n\} = \sum_{i=1}^{n-1} X_i$, in which $Y \sim \text{Bin} \left(n, \frac{p}{\lambda+p} \right)$, using the reasoning before Equation 4.4 and $X_i \sim \text{Geo} \left(1 - \frac{1-p}{\lambda+1} \right)$ by Lemma 4.1.2. Using this in Equation 4.9, we find

$$\begin{aligned}
\text{var} \left(\frac{|E|}{|V|} \mid |V| = n \right) &= \frac{n \cdot \frac{p}{\lambda+p} \cdot \frac{\lambda}{\lambda+p} + (n-1) \cdot \text{var} (X_i)}{n^2}, \\
&= \frac{n \cdot \frac{p}{\lambda+p} \cdot \frac{\lambda}{\lambda+p} + (n-1) \cdot \frac{\frac{1-p}{\lambda+1}}{\left(1 - \frac{1-p}{\lambda+1}\right)^2}}{n^2}, \\
&= \frac{n \cdot \frac{p}{\lambda+p} \cdot \frac{\lambda}{\lambda+p} + (n-1) \cdot \frac{\frac{1-p}{\lambda+1}}{\left(\frac{\lambda+p}{\lambda+1}\right)^2}}{n^2}, \\
&= \frac{n \cdot p \cdot \lambda + (n-1)(1-p)(\lambda+1)}{n^2 \cdot (\lambda+p)^2}. \tag{4.10}
\end{aligned}$$

□

Using this formula, we find that $\lim_{n \rightarrow \infty} \text{var} \left(\frac{|E|}{|V|} \mid |V| = n \right) = 0$ and it converges with rate $\frac{1}{n}$.

A different aspect we discuss is the expected number of edges per message tree in G_t . Again, we derive the expectation and the variance in Theorems 4.1.5 and 4.1.6 respectively.

Theorem 4.1.5 (Expected number of edges per message tree). *We consider the graph process at time t and derive the following expression for the expected number of edges per message tree (root-node of a superstar progression)*

$$\mathbb{E} \left[\frac{|E|}{|T|} \right] = \lambda^{-1} \cdot \left(1 - \left(\frac{1}{\lambda+1} \right)^t \right), \tag{4.11}$$

where $|T|$ denotes the number of message trees in the graph G_t .

Proof. Since all discussions that arrive are T1 arrivals and all retweets that arrive are T2 and T3 arrivals, we use Lemma 4.1.1 and find that the number of T2 and T3 arrivals follows a binomial distribution with parameters t and $\mathbb{P}(\text{T2 arrival}) + \mathbb{P}(\text{T3 arrival}) = \frac{1}{\lambda+1}$. When the amount of arrivals of T2 and T3 in G_t are known, the number of arrivals of T1 is also known (namely t minus the number of T2 and T3 arrivals). The amount of message trees is the number of arrivals of T1 plus one for the message tree in

G_0 . Using this we find

$$\begin{aligned}
\mathbb{E} \left[\frac{|E|}{|T|} \right] &= \mathbb{E} \left[\frac{\# \text{ T2 and T3 arrivals on } [0, t]}{\# \text{ T1 arrivals on } [0, t] + 1} \right], \\
&= \sum_{i=1}^t \frac{i}{t-i+1} \cdot \binom{t}{i} \cdot \left(\frac{1}{\lambda+1} \right)^i \cdot \left(\frac{\lambda}{\lambda+1} \right)^{t-i}, \\
&= \sum_{i=1}^t \frac{i}{t-i+1} \cdot \frac{t!}{i!(t-i)!} \cdot \left(\frac{1}{\lambda+1} \right)^i \cdot \left(\frac{\lambda}{\lambda+1} \right)^{t-i}, \\
&= \sum_{i=1}^t \frac{t!}{(i-1)!(t-i+1)!} \cdot \left(\frac{1}{\lambda+1} \right)^{i-1} \cdot \frac{1}{\lambda+1} \cdot \left(\frac{\lambda}{\lambda+1} \right)^{t-i+1} \cdot \left(\frac{\lambda}{\lambda+1} \right)^{-1}, \\
&= \frac{1}{\lambda+1} \cdot \left(\frac{\lambda}{\lambda+1} \right)^{-1} \cdot \sum_{i=1}^t \binom{t}{i-1} \cdot \left(\frac{1}{\lambda+1} \right)^{i-1} \cdot \left(\frac{\lambda}{\lambda+1} \right)^{t-i+1}, \\
&= \frac{1}{\lambda} \cdot \sum_{j=0}^{t-1} \binom{t}{j} \cdot \left(\frac{1}{\lambda+1} \right)^j \cdot \left(\frac{\lambda}{\lambda+1} \right)^{t-j}, \\
&= \frac{1}{\lambda} \cdot \left(1 - \left(\frac{1}{\lambda+1} \right)^t \right), \tag{4.12}
\end{aligned}$$

which completes the proof. Furthermore as $t \rightarrow \infty$, $\left(\frac{1}{\lambda+1} \right)^t \rightarrow 0$. Therefore

$$\lim_{t \rightarrow \infty} \mathbb{E} \left[\frac{|E|}{|T|} \right] = \lambda^{-1}. \tag{4.13}$$

□

As we shall see in Chapter 5, this expression is very useful for the parameter estimation. Using a similar argument, we can also derive the variance of the number of edges per message tree.

Theorem 4.1.6 (Variance for number of edges per message tree). *Given G_t we find that the variance for the number of edges per message tree is,*

$$\text{var} \left(\frac{|E|}{|T|} \right) = \frac{1}{\lambda} \left(\sum_{i=1}^t \frac{i}{t-i+1} \cdot \binom{t}{i-1} \left(\frac{1}{\lambda+1} \right)^{i-1} \left(\frac{\lambda}{\lambda+1} \right)^{t-i+1} - \frac{1}{\lambda} \cdot \left(1 - \left(\frac{1}{\lambda+1} \right)^t \right)^2 \right). \tag{4.14}$$

Proof. We obtain this equation as follows

$$\begin{aligned}
\text{var} \left(\frac{|E|}{|T|} \right) &= \text{var} \left(\frac{\# \text{ T2} + \# \text{ T3 on } [0, t]}{\# \text{ T1 on } [0, t]} \right), \\
&= \mathbb{E} \left[\left(\frac{\# \text{ T2} + \# \text{ T3 on } [0, t]}{\# \text{ T1 on } [0, t]} \right)^2 \right] - \mathbb{E} \left[\frac{\# \text{ T2} + \# \text{ T3 on } [0, t]}{\# \text{ T1 on } [0, t]} \right]^2. \tag{4.15}
\end{aligned}$$

In which $\mathbb{E} \left[\frac{\# \text{ T2} + \# \text{ T3 on } [0, t]}{\# \text{ T1 on } [0, t]} \right]$ is already known by Theorem 4.1.5. Therefore, we only need to derive an expression for the first term. We also use the proof of Theorem 4.1.5 as an outline to derive

$\mathbb{E} \left[\left(\frac{\# \text{T2} + \# \text{T3 on } [0, t]}{\# \text{T1 on } [0, t]} \right)^2 \right]$, in which we also use Lemma 4.1.1.

$$\begin{aligned}
\mathbb{E} \left[\left(\frac{\# \text{T2} + \# \text{T3 on } [0, t]}{\# \text{T1 on } [0, t]} \right)^2 \right] &= \sum_{i=0}^t \left(\frac{i}{t-i+1} \right)^2 \cdot \mathbb{P}(i \text{ arrivals of T2 or T3 on } [0, t]), \\
&= \sum_{i=0}^t \left(\frac{i}{t-i+1} \right)^2 \cdot \binom{t}{i} \left(\frac{1}{\lambda+1} \right)^i \left(\frac{\lambda}{\lambda+1} \right)^{t-i}, \\
&= \sum_{i=1}^t \left(\frac{i}{t-i+1} \right)^2 \cdot \frac{t!}{i!(t-i)!} \cdot \left(\frac{1}{\lambda+1} \right)^i \cdot \left(\frac{\lambda}{\lambda+1} \right)^{t-i}, \\
&= \frac{1}{\lambda+1} \cdot \left(\frac{\lambda}{\lambda+1} \right)^{-1} \cdot \sum_{i=1}^t \frac{i}{t-i+1} \cdot \binom{t}{i-1} \cdot \left(\frac{1}{\lambda+1} \right)^{i-1} \cdot \left(\frac{\lambda}{\lambda+1} \right)^{t-i+1}, \\
&= \frac{1}{\lambda} \cdot \sum_{i=1}^t \frac{i}{t-i+1} \cdot \binom{t}{i-1} \cdot \left(\frac{1}{\lambda+1} \right)^{i-1} \cdot \left(\frac{\lambda}{\lambda+1} \right)^{t-i+1}. \tag{4.16}
\end{aligned}$$

Combining this with Theorem 4.1.5 and Equation 4.15 we find

$$\begin{aligned}
\text{var} \left(\frac{|E|}{|T|} \right) &= \mathbb{E} \left[\left(\frac{\# \text{T2} + \# \text{T3 on } [0, t]}{\# \text{T1 on } [0, t]} \right)^2 \right] - \mathbb{E} \left[\frac{\# \text{T2} + \# \text{T3 on } [0, t]}{\# \text{T1 on } [0, t]} \right]^2, \\
&= \frac{1}{\lambda} \cdot \sum_{i=1}^t \frac{i}{t-i+1} \cdot \binom{t}{i-1} \cdot \left(\frac{1}{\lambda+1} \right)^{i-1} \cdot \left(\frac{\lambda}{\lambda+1} \right)^{t-i+1} - \left(\frac{1}{\lambda} \right)^2 \cdot \left(1 - \left(\frac{1}{\lambda+1} \right)^t \right)^2, \\
&= \frac{1}{\lambda} \left(\sum_{i=1}^t \frac{i}{t-i+1} \cdot \binom{t}{i-1} \cdot \left(\frac{1}{\lambda+1} \right)^{i-1} \cdot \left(\frac{\lambda}{\lambda+1} \right)^{t-i+1} - \frac{1}{\lambda} \cdot \left(1 - \left(\frac{1}{\lambda+1} \right)^t \right)^2 \right).
\end{aligned}$$

□

Since Equation 4.14 does not provide a good insight in the rate of the convergence of $\text{var} \left(\frac{|E|}{|T|} \right)$, we derive this rate of convergence as follows. First we rewrite the summation in Equation 4.14.

$$\begin{aligned}
&\sum_{i=1}^t \frac{i}{t-i+1} \cdot \binom{t}{i-1} \cdot \left(\frac{1}{\lambda+1} \right)^{i-1} \cdot \left(\frac{\lambda}{\lambda+1} \right)^{t-i+1}, \\
&= \sum_{i=1}^t \left(\frac{t+1}{t-i+1} - 1 \right) \cdot \binom{t}{i-1} \cdot \left(\frac{1}{\lambda+1} \right)^{i-1} \cdot \left(\frac{\lambda}{\lambda+1} \right)^{t-i+1}, \\
&= \sum_{i=1}^t \frac{t+1}{t-i+1} \cdot \binom{t}{i-1} \cdot \left(\frac{1}{\lambda+1} \right)^{i-1} \cdot \left(\frac{\lambda}{\lambda+1} \right)^{t-i+1} - \sum_{j=0}^{t-1} \binom{t}{j} \cdot \left(\frac{1}{\lambda+1} \right)^j \cdot \left(\frac{\lambda}{\lambda+1} \right)^{t-j}, \\
&= \mathbb{E} \left[\frac{t+1}{t-\tilde{X}+1} \right] - 1 + \left(\frac{1}{\lambda+1} \right)^t.
\end{aligned}$$

Where $\tilde{X} \sim \text{Bin}\left(t, \frac{1}{\lambda+1}\right)$. We then standardize the expectation in this last equation,

$$\begin{aligned} \mathbb{E}\left[\frac{t+1}{t-\tilde{X}+1}\right] &= \mathbb{E}\left[\frac{t+1}{\left(\frac{t}{\lambda+1} - \tilde{X}\right) + \frac{\lambda \cdot t + \lambda + 1}{\lambda+1}}\right], \\ &= \mathbb{E}\left[\frac{\frac{(t+1)(\lambda+1)}{\sqrt{\lambda t}}}{\frac{\lambda+1}{\sqrt{\lambda t}} \cdot \left(\frac{t}{\lambda+1} - \tilde{X}\right) + \frac{\lambda \cdot t + \lambda + 1}{\sqrt{\lambda t}}}\right], \\ &= \mathbb{E}\left[\frac{\frac{(t+1)(\lambda+1)}{\sqrt{\lambda t}}}{\tilde{Z} + \frac{\lambda \cdot t + \lambda + 1}{\sqrt{\lambda t}}}\right], \\ &= \frac{(t+1)(\lambda+1)}{\lambda t + \lambda + 1} \cdot \mathbb{E}\left[\frac{1}{\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot \tilde{Z} + 1}\right]. \end{aligned}$$

Within the function $\frac{1}{\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot \tilde{Z} + 1}$ we define $a(\lambda, t, \tilde{Z}) = \tilde{Z} \cdot \frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1}$. In this expectation, we distinguish two possibilities, namely $|a| > 1$ and $|a| < 1$. We then split the equation into two parts as follows:

$$\mathbb{E}\left[\frac{1}{\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot \tilde{Z} + 1}\right] = \mathbb{E}\left[\mathbb{1}_{|a| \geq 1} \left(\frac{1}{\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot \tilde{Z} + 1}\right)\right] + \mathbb{E}\left[\mathbb{1}_{|a| < 1} \left(\frac{1}{\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot \tilde{Z} + 1}\right)\right].$$

The first term of this expression goes to 0. Using the second term, we argue that $\text{var}\left(\frac{|E|}{|T|}\right)$ goes to 0 as $t \rightarrow \infty$ at rate $\frac{1}{t}$. By the central limit theorem (CLT) we know that $\tilde{Z} \xrightarrow{D} Z \sim N(0, 1)$ for $t \rightarrow \infty$. Therefore we replace \tilde{Z} with Z . Furthermore, we omit $\mathbb{1}_{|a| < 1}$. We then take the Maclaurin Series,

$$\mathbb{E}\left[\frac{1}{\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot Z + 1}\right] = \mathbb{E}\left[1 - \frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot Z + \left(\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot Z\right)^2 - \left(\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot Z\right)^3 + \dots\right].$$

We rewrite this equation using

$$-\mathbb{E}\left[\left|\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot Z\right|^3\right] \leq \mathbb{E}\left[-\left(\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot Z\right)^3 + \left(\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot Z\right)^4 - \dots\right] \leq \mathbb{E}\left[\left|\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot Z\right|^3\right].$$

Furthermore, using Hall [21], we find that that $\mathbb{E}\left[|\tilde{Z}|^p\right] \rightarrow \mathbb{E}[|Z|^p]$ for $p > 0$ under the CLT, we find

$$\mathbb{E}\left[\left|\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot Z\right|^3\right] = \left|\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1}\right|^3 \cdot 2\sqrt{\frac{2}{\pi}}.$$

Using these equations, we derive bounds for $\mathbb{E}\left[\frac{1}{\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot Z + 1}\right]$,

$$\begin{aligned} &1 + \mathbb{E}\left[\left(\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot Z\right)^2\right] - \left|\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1}\right|^3 \cdot 2\sqrt{\frac{2}{\pi}}, \\ &\leq \mathbb{E}\left[\mathbb{1}_{|a| < 1} \left(\frac{1}{\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot Z + 1}\right)\right], \\ &\leq 1 + \mathbb{E}\left[\left(\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot Z\right)^2\right] + \left|\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1}\right|^3 \cdot 2\sqrt{\frac{2}{\pi}}, \\ &= 1 + \frac{\lambda t}{(\lambda t + \lambda + 1)^2} + \left|\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1}\right|^3 \cdot 2\sqrt{\frac{2}{\pi}} \end{aligned}$$

We use these equations to find the limit of the variance for $t \rightarrow \infty$,

$$\begin{aligned}
\lim_{t \rightarrow \infty} \text{var} \left(\frac{|E|}{|T|} \right) &= \lim_{t \rightarrow \infty} \frac{1}{\lambda} \left(\mathbb{E} \left[\frac{t+1}{t-\tilde{X}+1} \right] - 1 + \left(\frac{1}{\lambda+1} \right)^t - \frac{1}{\lambda} \cdot \left(1 - \left(\frac{1}{\lambda+1} \right)^t \right)^2 \right), \\
&\approx \lim_{t \rightarrow \infty} \frac{1}{\lambda} \left(\frac{(t+1)(\lambda+1)}{\lambda t + \lambda + 1} \cdot \mathbb{E} \left[\frac{1}{\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot Z + 1} \right] - 1 + \left(\frac{1}{\lambda+1} \right)^t - \frac{1}{\lambda} \cdot \left(1 - \left(\frac{1}{\lambda+1} \right)^t \right)^2 \right), \\
&= \lim_{t \rightarrow \infty} \frac{1}{\lambda} \left(\frac{(t+1)(\lambda+1)}{\lambda t + \lambda + 1} \cdot \left(1 + \frac{\lambda t}{(\lambda t + \lambda + 1)^2} + \left| \frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \right|^3 \cdot 2\sqrt{\frac{2}{\pi}} \right) - 1 + \left(\frac{1}{\lambda+1} \right)^t - \frac{1}{\lambda} \cdot \left(1 - \left(\frac{1}{\lambda+1} \right)^t \right)^2 \right), \\
&= \frac{1}{\lambda} \left(\frac{\lambda+1}{\lambda} (1+0+0) - 1 + 0 - \frac{1}{\lambda} (1-0) \right), \\
&= \frac{1}{\lambda} \left(\frac{\lambda+1}{\lambda} - \frac{\lambda+1}{\lambda} \right) = 0.
\end{aligned}$$

We also find that $\mathbb{E} \left[\frac{t+1}{t-\tilde{X}+1} \right]$ converges with rate $\frac{1}{t}$ since,

$$\begin{aligned}
&\lim_{t \rightarrow \infty} t \cdot \frac{(t+1)(\lambda+1)}{\lambda \cdot t + \lambda + 1} \cdot \left(\mathbb{E} \left[\frac{1}{\frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \cdot Z + 1} \right] - 1 \right) \\
&= \lim_{t \rightarrow \infty} t \cdot \frac{(t+1)(\lambda+1)}{(\lambda t + \lambda + 1)} \cdot \left(\frac{\lambda t}{(\lambda t + \lambda + 1)^2} + \left| \frac{\sqrt{\lambda t}}{\lambda t + \lambda + 1} \right|^3 \cdot 2\sqrt{\frac{2}{\pi}} \right), \\
&= \frac{(\lambda+1)}{\lambda^2},
\end{aligned}$$

is a constant. And since this expectation in the equation converges with rate $\frac{1}{t}$, so does the variance.

4.2 Component size distribution

In this section we assume that G_t consists of m connected components (C_1, C_2, \dots, C_m) with known respective sizes $(|C_1|, \dots, |C_m|)$. We aim to derive expressions for the distribution of the component sizes after the first arrival to G_t . To do this, we first derive the probabilities that components merge.

Lemma 4.2.1. *Given a graph G_t , we can derive the probability that components C_i and C_j merge given that a T3 arrival occurs next.*

$$\mathbb{P}(\text{components } j \text{ and } i \text{ join} \mid \text{T3 arrival}) = \frac{2 \cdot |C_i| \cdot |C_j|}{|V|^2 - |V|}. \quad (4.17)$$

Furthermore it holds that

$$\mathbb{P}(\text{components merge} \mid \text{T3 arrival}) = \frac{2 \cdot \sum_k |C_k| \cdot \sum_{l>k} |C_l|}{|V|^2 - |V|} \quad (4.18)$$

Proof. In this proof all probabilities are conditioned on the event that there is a T3 arrival. We omit this conditioning for brevity. Consider all components (in short comp.) to be urns and all nodes in the components as balls in the corresponding urn. Using this interpretation we can express the desired probability as the amount of outcomes when selecting two balls from two different urns, divided by the

total number of options. When C_i and C_j merge, we have $|C_i| \cdot |C_j|$ possible outcomes to get the desired balls from urns i and j . All possible combinations of selecting two balls from all urns are

$$\binom{|V|}{2} = \frac{|V|!}{2! \cdot (|V| - 2)!} = \frac{|V| \cdot (|V| - 1)}{2} = \frac{|V|^2 - |V|}{2}.$$

Thus we obtain to Equation 4.17. Moreover, we can join any two components C_k and C_l . The total amount of possibilities for this event to occur is $\sum_k |C_k| \cdot \sum_{l>k} |C_l|$. Therefore, we also find the following probability

$$\mathbb{P}(\text{components } j \text{ and } i \text{ join} \mid \text{components merge}) = \frac{|C_i| \cdot |C_j|}{\sum_k |C_k| \cdot \sum_{l>k} |C_l|}.$$

Then by using the multiplication formula, we obtain

$$\mathbb{P}(\text{comp. } j \text{ and } i \text{ join} \mid \text{comp. merge}) \cdot \mathbb{P}(\text{comp. merge}) = \mathbb{P}(\text{comp. merge, comp. } j \text{ and } i \text{ join}). \quad (4.19)$$

Using this expression, combined with the two probabilities found earlier, we find the probability that two components of G_t merge on the next arrival.

$$\begin{aligned} \mathbb{P}(\text{components merge}) &= \frac{\mathbb{P}(\text{components merge, components } j \text{ and } i \text{ join})}{\mathbb{P}(\text{components } j \text{ and } i \text{ join} \mid \text{components merge})} \\ &= \frac{\frac{2 \cdot |C_i| \cdot |C_j|}{|V|^2 - |V|}}{\frac{|C_i| \cdot |C_j|}{\sum_k |C_k| \cdot \sum_{l>k} |C_l|}} \\ &= \frac{2 \cdot \sum_k |C_k| \cdot \sum_{l>k} |C_l|}{|V|^2 - |V|} \end{aligned} \quad (4.20)$$

This completes the lemma. \square

We use these probabilities to derive expressions for the distribution of the sizes of the components of G_{t+1} .

Lemma 4.2.2. *The distribution of the sizes of the components of G_{t+1} , given G_t is as follows,*

$$\text{component sizes of } G_{t+1} : \begin{cases} |C_1|, \dots, |C_i|, |C_j|, \dots, |C_m|, 1 & \text{w.p. } \frac{\lambda}{\lambda+1} \\ |C_1|, \dots, |C_i| + 1, |C_j|, \dots, |C_m| & \text{w.p. } \frac{p}{\lambda+1} \cdot \frac{|C_i|}{|V|} \\ |C_1|, \dots, |C_i| + |C_j|, \dots, |C_m| & \text{w.p. } \frac{1-p}{\lambda+1} \cdot \frac{2 \cdot |C_i| \cdot |C_j|}{|V|^2 - |V|} \\ |C_1|, \dots, |C_i|, |C_j|, \dots, |C_m| & \text{w.p. } \frac{1-p}{\lambda+1} \cdot \frac{|C_i|^2 - |C_i|}{|V|^2 - |V|} \end{cases} \quad (4.21)$$

Proof. There are three types of arrivals, which we will discuss separately. First at a T1 arrival, which occurs w.p. $\frac{\lambda}{\lambda+1}$, we create a new component, thus the new distribution of component sizes is

$$G_{t+1} = \{C_1, \dots, C_m, C_{m+1}\},$$

in which $|C_{m+1}| = 1$.

Then we consider a T2 arrival, which occurs w.p. $\frac{p}{\lambda+1}$. We now add a node to an existing component C_i w.p. $\frac{|C_i|}{|V|}$. Thus the probability that we add the new node to C_i is $\frac{p}{\lambda+1} \cdot \frac{|C_i|}{|V|}$.

Last, we consider a T3 arrival. In this case we have two options. The new edge can either join two components, or join two nodes that are already in one component. For the first case, we take Equation 4.17 from Lemma 4.2.1 to derive the probability that C_i and C_j join as

$$\mathbb{P}(\text{Components } C_i \text{ and } C_j \text{ merge}) = \mathbb{P}(\text{T3 arrival}) \cdot \frac{2 \cdot |C_i| \cdot |C_j|}{|V|^2 - |V|} = \frac{1-p}{\lambda+1} \cdot \frac{2 \cdot |C_i| \cdot |C_j|}{|V|^2 - |V|}.$$

Then for the second case, we use a similar reasoning as the proofs in Lemma 4.2.1. The number of ways a T3 arrival links two nodes that are already connected in a component, say C_i , is $|C_i|(|C_i| - 1)$. Therefore with probability $\frac{|C_i|^2 - |C_i|}{|V|^2 - |V|}$ the component size does not change.

Finally we check if these last two cases cover all possibilities. The second term in our summation is equal to the summation in Equation 4.20. Therefore, we simplify the following equation:

$$\begin{aligned}
\sum_i [|C_i|^2 - |C_i|] + 2 \cdot \sum_i |C_i| \sum_{j>i} |C_j| &= \sum_i |C_i|^2 - |V| + 2 \cdot \sum_i |C_i| \sum_{j>i} |C_j| \\
&= \sum_i \sum_j |C_i| |C_j| - |V| \\
&= |V|^2 - |V|
\end{aligned} \tag{4.22}$$

□

4.2.1 The case $p=1$: a Pólya process

So far we attempted to derive expressions for the distribution of the component sizes for G_t , given the current distribution. Another option is to derive the distribution of the component size as a whole. For this analysis we need the notion of exchangeability.

Definition 4.2.3 (exchangeable random variables). Let X_1, X_2, \dots, X_k be a sequence of random variables. This sequence is called exchangeable if and only if for every permutation σ of $1, 2, \dots, k$ it holds that,

$$\mathbb{P}(X_{\sigma(1)} = x_{\sigma(1)}, X_{\sigma(2)} = x_{\sigma(2)}, \dots, X_{\sigma(k)} = x_{\sigma(k)}) = \mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k). \tag{4.23}$$

Thus the joint probability distribution of the sequence is the same, regardless of the order of the events.

If we then assume that $p = 1$ in our model, we can use the infinite generalized Pólya process (analysed by Chung et al. in [15]), which uses two parameters, γ and \bar{p} . It is defined as follows,

Definition 4.2.4 (infinite generalized Pólya process). For fixed parameters $\gamma \in \mathbb{R}, 0 \leq \bar{p} < 1$, begin with a bin containing one ball and then introduce balls one at a time. For each new ball, with probability \bar{p} , create a new bin and place the ball in that bin; with probability $1 - \bar{p}$, place the ball in an existing bin, such that the probability that the ball is placed in a bin is proportional to x^γ , where x is the number of balls in that bin.

Also, we define $f_i \propto g(i)$ as $f_i = c(1 + o(1))g(i)$ for some constant c . Let f_i denote the fraction of bins that contain i balls, then

$$f_i \propto i^{-(1+\frac{1}{1-\bar{p}})}. \tag{4.24}$$

Equation 4.24 is derived in [15] by first calculating a recurrence relation between f_i and f_{i-1} . This recurrence is then used to determine Equation 4.24.

Returning to our model, we take $p = 1$. In this situation message trees in our model cannot be combined. Therefore we only have arrivals of T1 and T2. Consider every message tree as a bin and every user inside the message tree as a ball. Adding a new ball to a bin then is equivalent to adding a retweet to a message tree and adding a new bin with a ball in it is equivalent to adding a new message tree to the progression. Since the probability that a particular bin is chosen depends on the number of balls in that bin, we choose $\gamma = 1$ and the probability of an addition of a new bin is $\bar{p} = \mathbb{P}(\text{T1 arrival}) = \frac{\lambda}{\lambda+1}$. Thus our model is equivalent to an infinite generalized Pólya process with parameters $\gamma = 1$ and $\bar{p} = \frac{\lambda}{\lambda+1}$ if $p = 1$. Then, by using Equation 4.24 and the fact that $\bar{p} = \frac{\lambda}{\lambda+1}$, we find

$$f_i \propto i^{-(2+\lambda)} \tag{4.25}$$

Thus the component size distribution follows a power-law distribution with parameter $2 + \lambda$ when $p = 1$. We can compare these results to simulations using the model. The measure that we use in this comparison is the complementary cumulative density function (CCDF) of the component size distribution, defined as

$$\overline{F(x)} = \mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x) = 1 - c \cdot \sum_{i=0}^x f_i. \tag{4.26}$$

In which c is a constant such that $c \cdot \sum_{i=0}^{\infty} f_i = 1$. We execute 5 runs with $\lambda = \frac{3}{4}$, $p = 1$, $q = 0.95$. In these runs we use 3 stop criteria for n , being $n = 1000$, $n = 10000$ and $n = 100000$. Notice that the parameter q does not influence the component size. The results of these runs are depicted in Figure 4.1. In this Figure, we use the black line to indicate the value of the CCDF using Equation 4.26. We see that as n increases, the outcome of the runs follows a line with the same slope as the black line, which supports our statement. The lines do not overlap exactly because we used a normalized form of f_i in Figure 4.1, whereas Equation 4.25 states that f_i is proportional to that function.

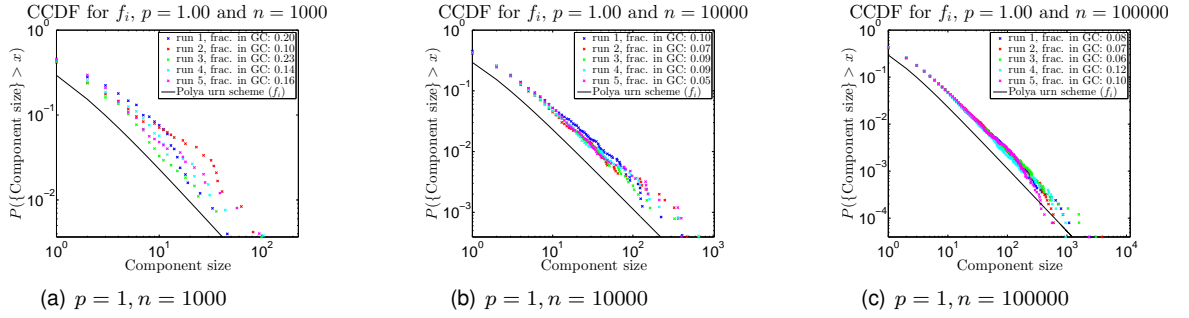


Figure 4.1: Complementary Cumulative Density Function for f_i .

4.3 Size of the giant component

The ultimate goal of our analysis is to derive an expression for $\frac{|E_{GC}|}{|V_{GC}|}$. Through Lemma 4.2.2 we can derive the probability distribution for the size of the largest component for a small graph. Below, we derive the possible outcomes for G_3 and use them to form the probability distribution of the giant component size. We find that,

$$G_3 \sim \begin{cases} \{1\} & \text{w.p. } \frac{(1-p)^3}{(\lambda+1)^3}, \\ \{2\} & \text{w.p. } \frac{3p(1-p)^2 + 2\lambda(1-p)^2}{(\lambda+1)^3}, \\ \{3\} & \text{w.p. } \frac{3p^2(1-p) + \frac{7}{3}\lambda p(1-p)}{(\lambda+1)^3}, \\ \{4\} & \text{w.p. } \frac{p^3}{(\lambda+1)^3}, \\ \{1, 1\} & \text{w.p. } \frac{\lambda(1-p)^2}{(\lambda+1)^3}, \\ \{2, 1\} & \text{w.p. } \frac{\frac{11}{3}\lambda p(1-p) + 2\lambda^2(1-p)}{(\lambda+1)^3}, \\ \{2, 2\} & \text{w.p. } \frac{\frac{2}{3}\lambda p^2}{(\lambda+1)^3}, \\ \{3, 1\} & \text{w.p. } \frac{\frac{7}{3}\lambda p^2}{(\lambda+1)^3}, \\ \{1, 1, 1\} & \text{w.p. } \frac{(1-p)\lambda^2}{(\lambda+1)^3}, \\ \{2, 1, 1\} & \text{w.p. } \frac{3\lambda^2 p}{(\lambda+1)^3}, \\ \{1, 1, 1, 1\} & \text{w.p. } \frac{\lambda^3}{(\lambda+1)^3}. \end{cases}$$

And thus, we can calculate the probability distribution of the giant component size for G_3 :

$$|V_{GC}| \mid G_3 \sim \begin{cases} 1 & \text{w.p. } \frac{(1-p)^3 + \lambda(1-p)^2 + (1-p)\lambda^2 + \lambda^3}{(\lambda+1)^3}, \\ 2 & \text{w.p. } \frac{3p(1-p)^2 + 2\lambda(1-p)^2 + \frac{11}{3}\lambda p(1-p) + 2\lambda^2(1-p) + \frac{2}{3}\lambda p^2 + 3\lambda^2 p}{(\lambda+1)^3}, \\ 3 & \text{w.p. } \frac{3p^2(1-p) + \frac{7}{3}\lambda p(1-p) + \frac{7}{3}\lambda p^2}{(\lambda+1)^3}, \\ 4 & \text{w.p. } \frac{p^3}{(\lambda+1)^3}. \end{cases}$$

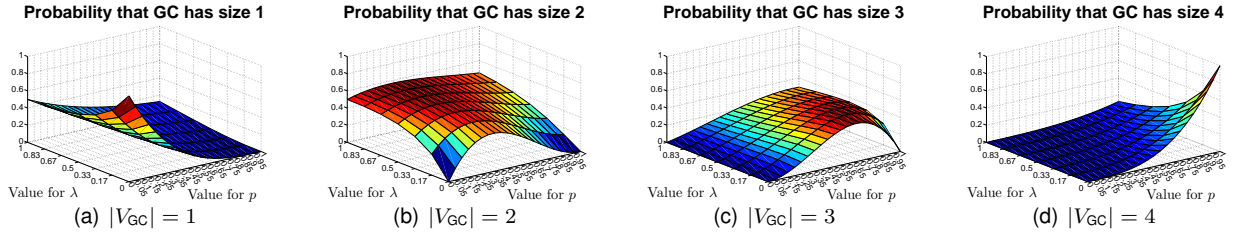


Figure 4.2: Probabilities for Giant Component in G_3 .

In Figure 4.2 we find four plots that indicate the probability that the giant component has a given size. On the x-axis and the y-axis we display several values for λ and p , the z-axis displays the probability that the giant component is of the size indicated in the caption of the figure. We clearly see that in Figure 4.2(a) that the giant component has size 1 for low values of p and that this probability increases as λ goes to 0. In Figure 4.2(d) we see that the giant component has size 4 if p is large and λ is low. Then Figure 4.2(c) indicates that the giant component is of size 3 when p is approximately 0.7 and λ is low. Finally, we see in Figure 4.2(b) that the giant component has size 2 if p is low for low λ values and that this probability becomes higher when λ increases. Thus, the giant component gets larger for low values of λ . The same holds for larger values for p .

Another strategy to derive an estimate for the size of the giant component is using the distribution of component sizes. In this estimate we assume that all T3 arrivals occur after all nodes have arrived to the graph. However, we cannot use this estimate as a bound for the size of the giant component. We show this by the following example. Consider two moments in time, s and t . We assume $s \leq t$. Let G_s contain two components i and j of size 1, thus $|C_i^s| = |C_j^s| = 1$. Furthermore we assume that $|V^s| = 5$ and $|V^t| = 20$. Using Lemma 4.2.1 we know that,

$$\mathbb{P}(i \text{ and } j \text{ merge at time } s) = \frac{2 \cdot |C_i^s| \cdot |C_j^s|}{|V^s|^2 - |V^s|} = \frac{2}{20}.$$

Similarly we know,

$$\mathbb{P}(i \text{ and } j \text{ merge at time } t) = \frac{2 \cdot |C_i^t| \cdot |C_j^t|}{|V^t|^2 - |V^t|} = \frac{2 \cdot |C_i^t| \cdot |C_j^t|}{380}.$$

Thus we conclude:

$$\begin{aligned} |C_i^t| \cdot |C_j^t| = 19 &\implies \mathbb{P}(i \text{ and } j \text{ merge at time } s) = \mathbb{P}(i \text{ and } j \text{ merge at time } t), \\ |C_i^t| \cdot |C_j^t| < 19 &\implies \mathbb{P}(i \text{ and } j \text{ merge at time } s) > \mathbb{P}(i \text{ and } j \text{ merge at time } t), \\ |C_i^t| \cdot |C_j^t| > 19 &\implies \mathbb{P}(i \text{ and } j \text{ merge at time } s) < \mathbb{P}(i \text{ and } j \text{ merge at time } t). \end{aligned}$$

Therefore, we can only use this approach to obtain an estimate of the size of the giant component, without knowing if it is larger or smaller than the actual value.

CHAPTER 5

THE MODEL IN PRACTICE

In this chapter we describe how we obtain estimates for the parameters from available tweets. Then, we use these expressions to perform a numerical verification of the expressions in Chapter 4. Subsequently, we analyse the sensitivity of the parameters to changes through simulations. Finally, we test the performance of the model with respect to the datasets that we discussed in Chapter 2.

5.1 Deriving parameter estimators

In this section we derive expressions which can be used to estimate the parameters λ , p and q of our model. First we derive an expression to estimate λ . Using Theorem 4.1.5, we know that $\frac{|E|}{|T|}$ goes to λ^{-1} as $t \rightarrow \infty$. Thus, we estimate λ as follows,

$$\hat{\lambda} = \frac{|T|}{|E|}. \quad (5.1)$$

Second, we derive an expression for \hat{p} . Since we already have an expression for $\hat{\lambda}$, we can combine this expression with Theorem 4.1.3 to obtain,

$$\begin{aligned} \frac{|E|}{|V|} &= \frac{1}{\hat{\lambda} + \hat{p}} - \frac{1 - \hat{p}}{|V| \cdot (\hat{\lambda} + \hat{p})} \\ (\hat{\lambda} + \hat{p}) \cdot |E| &= |V| - 1 + \hat{p} \\ \hat{p} \cdot |E| - \hat{p} &= |V| - \hat{\lambda} \cdot |E| - 1 \\ \hat{p} &= \frac{|V| - |T| - 1}{|E| - 1} \end{aligned} \quad (5.2)$$

Third, we derive an expression for \hat{q} . In this expression let E_T denote the set of edges that have a root node of a message tree as a source node. Then from the definition of the superstar model [8], we derive the estimator

$$\hat{q} = \frac{|E_T|}{|E|}. \quad (5.3)$$

Notice that we can obtain the needed numbers ($|E|$, $|T|$, $|V|$ and $|E_T|$) directly from a dataset of tweets. In Section 5.4 we estimate the parameters using the tweets that have been posted. Since more tweets arrive over time in our datasets, the parameter estimates vary over time.

5.2 Verification through simulation

In this section, we connect the output of our model to the theoretic values found in Chapter 4. For this analysis we set all parameters and compare the results of multiple runs with the equations we found. In this section, we start a simulation with G_0 , containing only one node.

We set $\lambda = \frac{1}{3}$, $p = \frac{3}{4}$ and $q = \frac{19}{20}$. Since Theorems 4.1.3 and 4.1.4 are conditioned on the number of nodes in the graph, we stop the simulation if $|V| = n$. Using Theorems 4.1.3 and 4.1.4, we find theoretical values for several values of n (indicated by the green solid and dotted lines in Figure 5.1(a)). In the same figure, we depict the average value (blue line) and its standard deviation (red dotted lines)

for 50 simulations of our model with the predefined parameters. These simulations are executed for multiple values of n .

Similarly, we use Theorems 4.1.5 and 4.1.6 with several values for t . Again, we set $\lambda = \frac{1}{3}$, $p = \frac{3}{4}$, $q = \frac{19}{20}$ and simulate 50 runs and depict the results in Figure 5.1(b). During these simulations, we used the number of time steps t as a stop criterion for the simulation. For values above $t = 1000$, we cannot numerically determine the values corresponding to Theorem 4.1.6, since $\binom{t}{k}$ gives too large numbers for large values of t .

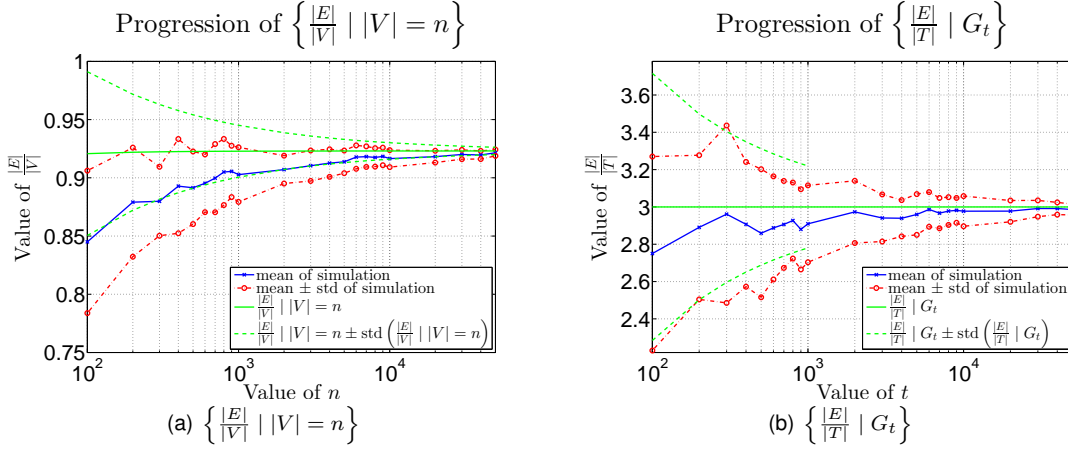


Figure 5.1: Theoretical and simulated values for 50 runs.

In Figure 5.1 we find that the simulated results and the theoretical results are consistent. For the simulation results of $\mathbb{E} \left[\frac{|E|}{|V|} \mid |V| = n \right]$ we find that they lie below the theoretical values, this can be explained by the following: In our simulation, when we try to join an edge to the graph, we check if that edge already exists, since we do not allow multiple edges. If the new edge already exists, we do nothing. Since we do not distinguish this in the theoretical case, there are more edges in this graph which gives a higher average degree. As time progresses, there are more possibilities for edges to join the graph, which induces a lower probability that we try to join an existing edge to the graph.

5.3 Sensitivity analysis

Another interesting aspect is to know if the output can be changed dramatically by a slight change in the parameters. With this analysis we test the robustness of the model and we gain insight in the relation between the parameters. We perform this analysis in this section by fixing two out of three parameters and executing multiple runs using various values for the third parameter. Again, we start the simulations with a graph G_0 , consisting of one node. For this analysis we execute multiple simulations using our model, again we perform 50 runs. The results that we display in this section are the average values of all individual runs for the given parameters.

The fixed values are $\lambda = \frac{1}{3}$, $p = 0.75$ and $q = 0.95$. For the analysis we take $\lambda \in \left\{ \frac{1}{3}, \frac{2}{3}, 1, \frac{4}{3}, \frac{5}{3}, 2 \right\}$, $p \in \{0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.99, 1\}$ and $q \in \{0.9, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.99, 1\}$. We compare the results on five measures. These measures are $\frac{|E|}{|V|}$, $\frac{|E_{GC}|}{|V_{GC}|}$, $\frac{|V_{GC}|}{|V|}$, $\frac{|E_{GC}|}{|E|}$ and the diameter.

In Figures 5.2 and 5.3 we compare the measures $\frac{|E|}{|V|}$ and $\frac{|E_{GC}|}{|V_{GC}|}$. First, in Figures 5.2(c) and 5.3(c), we find that the outcomes are similar for all different values of q . Then in Figure 5.2(b) we see that the values for $\frac{|E|}{|V|}$ range from 0.7 to 1.2, so there is a transition from a tree-like structure to a denser structure when p is lower than 0.65. Furthermore, the outcomes are consistent through time for high values of p . Also, the results for $p = 1$ are less than 1 for all t , as shown in Figure 5.3(b). In Figures 5.2(a) and 5.3(a) we see that different values for λ can also give a wide range of outcomes for the both measures. Here

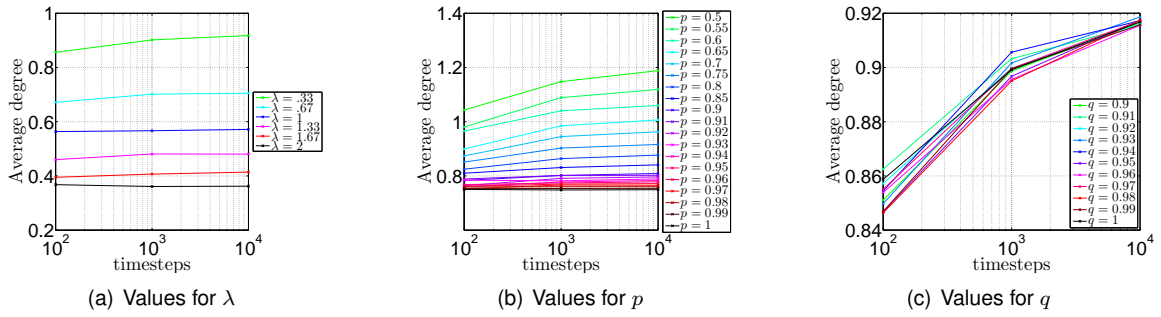


Figure 5.2: Analysis for measure $\frac{|E|}{|V|}$.

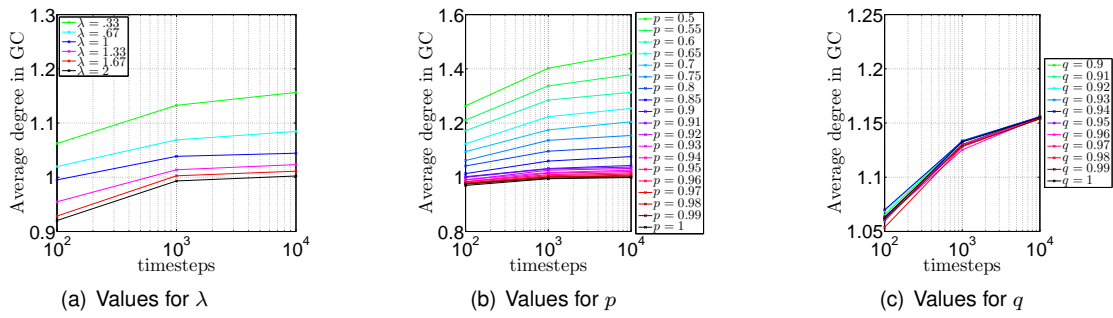


Figure 5.3: Analysis for measure $\frac{|E_{GC}|}{|V_{GC}|}$.

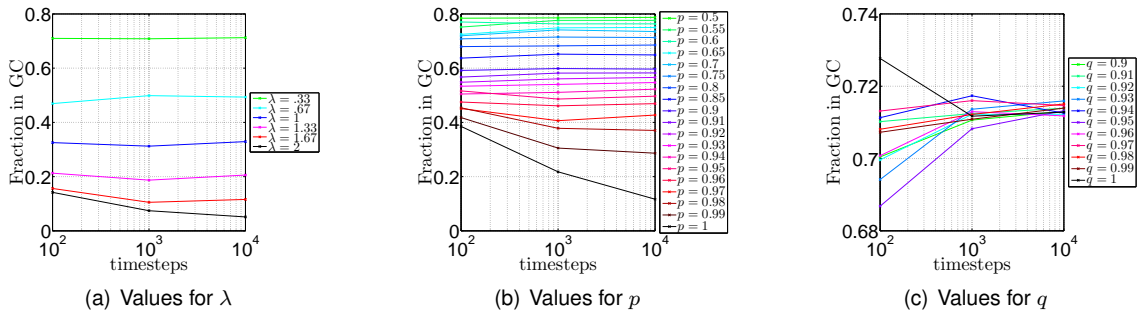


Figure 5.4: Analysis for measure node percentage in GC.

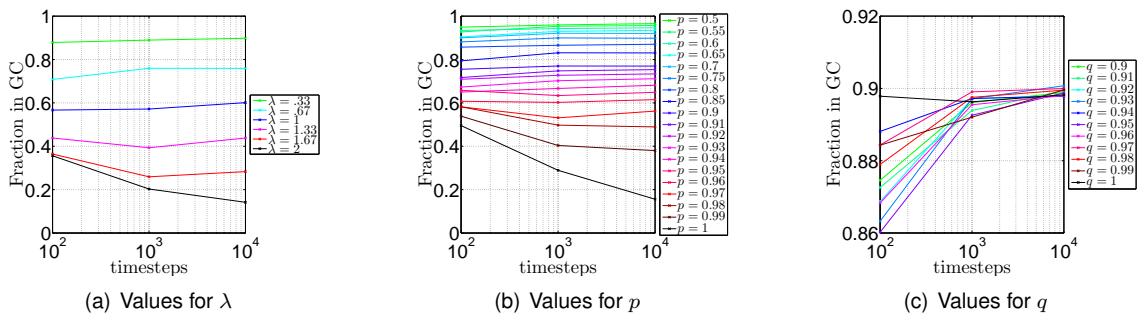


Figure 5.5: Analysis for measure edge percentage in GC.

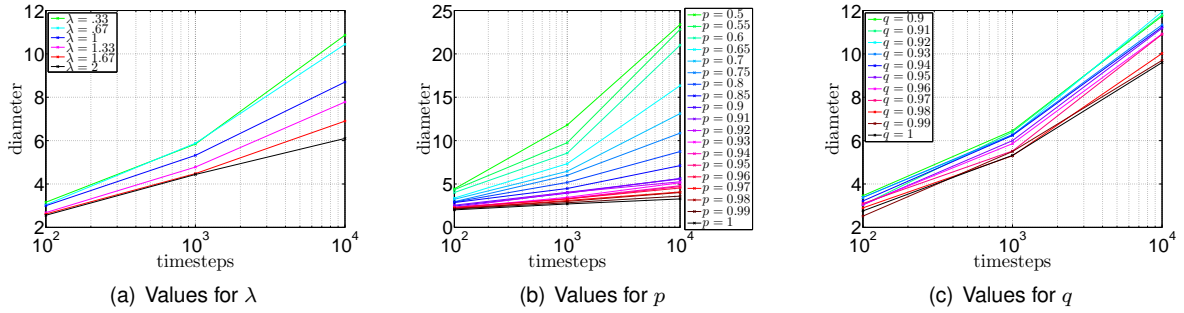


Figure 5.6: Analysis for diameter.

it also holds that for low λ , $\frac{|E|}{|V|}$ and $\frac{|E_{GC}|}{|V_{GC}|}$ are large. Thus we may conclude that these measures heavily depend on the values for p and λ .

The next two measures correspond to the size of the giant component of the graph, which are depicted in Figures 5.4 and 5.5. Again, we see that λ and p are the most decisive parameters. When p approaches 1, the fraction of nodes and edges in the GC drop significantly as t grows. Again for q the results are consistent for all values of q as t grows. However, the values are not differ considerably for lower values of t .

For the last measure, the diameter, we find that it grows over time for all parameters, regardless of their values. For q all outcomes are similar, see Figure 5.6(c). For p , we see that if p is smaller than 0.9 the diameter increases more rapidly as t grows, which is indicated in Figures 5.6(b). This also holds for λ when it is lower than 1 (Figure 5.6(a)).

5.4 Prediction using the estimates

For every hour in the progression of a dataset, we determine the tweets that were posted until this time. Using these tweets we estimate the parameters using Equations 5.1, 5.2 and 5.3. Then we use our model to build a graph to the same size as the end progression. We then calculate the properties of these predicted progressions. For every hour, we perform 50 of these predictions, of which we present the average results. Furthermore, we compare the properties of some simulations to the properties in the actual dataset. The observations and conclusions we state in this section are based on a combined figure of the estimates, the normalized number of tweets and the properties. Since these figures are hard to comprehend, we present the figures separately.

5.4.1 Project-X dataset

In Figure 5.7 we display the parameter estimates at certain moments using the tweets that are available up to that time. At the start of the progression, the estimate of λ decreases significantly. This means that there are more retweets arriving compared to new discussions. This time coincides with the forming of a large component. Lastly we see that the estimates of all parameters are very stable after the peak in the dataset, since the largest part of the graph is already visible at that time.

The results of the simulations are displayed in Figure 5.8. The most remarkable in Figure 5.8 is the decrease of the predicted values for $\frac{|V_{GC}|}{|V|}$ and $\frac{|E_{GC}|}{|V_{GC}|}$ during 22-9. This decrease coincides with the increase of \hat{p} that occurs just before the peak. This can be explained as follows: A higher value of \hat{p} causes less a lower probability of merging components in the simulation, which has a smaller expected giant component size as a consequence.

Although the predicted values of these measures underestimate the real values in the dataset (indicated by the black line in Figure 5.8), they already indicate that the giant component is going to be of a considerable size and is denser than a tree-like structure. The results that are displayed after 22-9 are almost identical to the actual graph, since the largest part of the graph is already visible at that time.

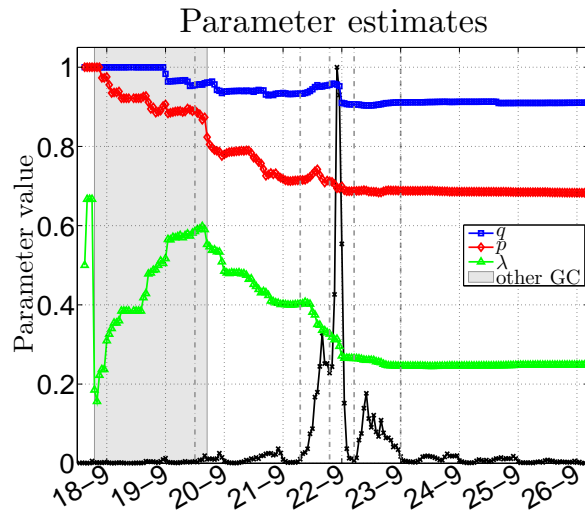


Figure 5.7: Progression of the parameter estimates using the tweets that are available until the indicated time.

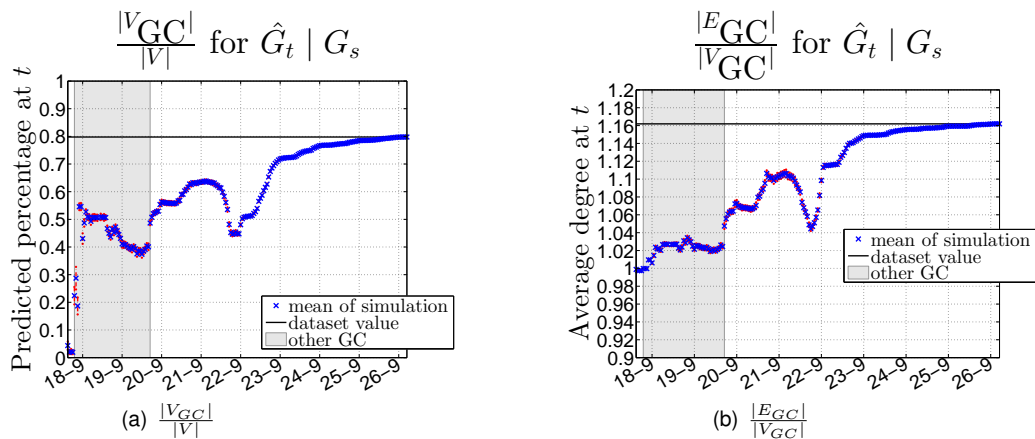


Figure 5.8: *Project-X* simulations using data that is available to that time (standard deviations are indicated in red).

We compare the result of the prediction of the graph with the actual graph for four different times, namely the early progression (19-9-2012 12:00), the densification (20-9-2012 0:00), the start of the peak (21-9-2012 7:00) and the middle of the peak (21-9-2012 19:00). First, we consider the simulated progression of the graph at 19-9-2012 12:00, displayed in 5.9. Due to the high value of \hat{p} , we see that the in-degree (Figure 5.9(a)) and the component-size distribution (Figure 5.9(c)) do not coincide at all the dataset values. If \hat{p} has a high value, components do not merge often, resulting in many small components. The out-degree (Figure 5.9(b)) appears to resemble the dataset well, with the exception of the outliers in the dataset.

Then, in the comparison of the dataset with the simulation for the densification at 2012-9-20 0:00 (see Figure 5.10), we find that the in-degree and the component size are a lot better with respect to the early progression simulation. Still, the in-degree in the simulation is significantly lower than in the actual dataset. The size of the giant component has increased significantly and is with its 60% a lot closer to the actual 80% in the dataset.

Next we display a comparison with the start of the peak (2012-9-21 7:00) in Figure 5.11. we find that only the component size distribution improves with respect to our last estimate. All other distributions perform less than at the densification.

Finally, at the simulation at the middle of the peak (2012-9-21 19:00), displayed in Figure 5.12, we see a decrease in the accuracy of all the distributions with respect to earlier simulations. This underlines the performance of the simulation seen in 5.8.

Thus we see that our model indicates that a large component is going to be present in the final progression the *Project-X* tweets. Also the out-degree distribution of the graph is similar for the simulations and the actual dataset. The predictive power of the in-degree distribution is a lot less, this is an aspect of the model that needs improvement.

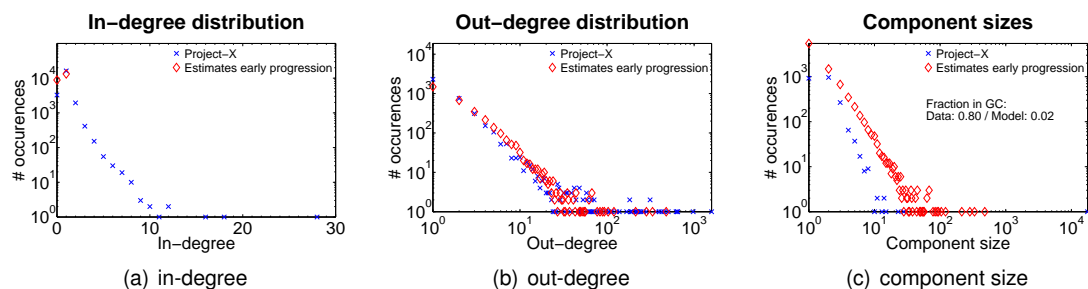


Figure 5.9: Comparison of *Project-X* and simulation using estimates from 2012-9-17 18:00

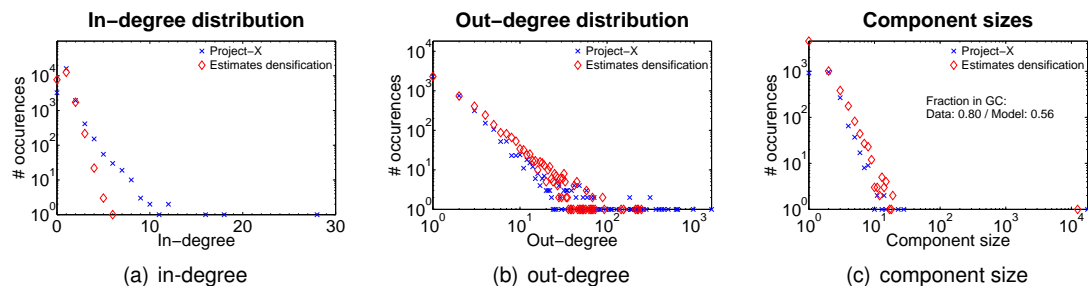


Figure 5.10: Comparison of *Project-X* and simulation using estimates from 2012-9-20 0:00

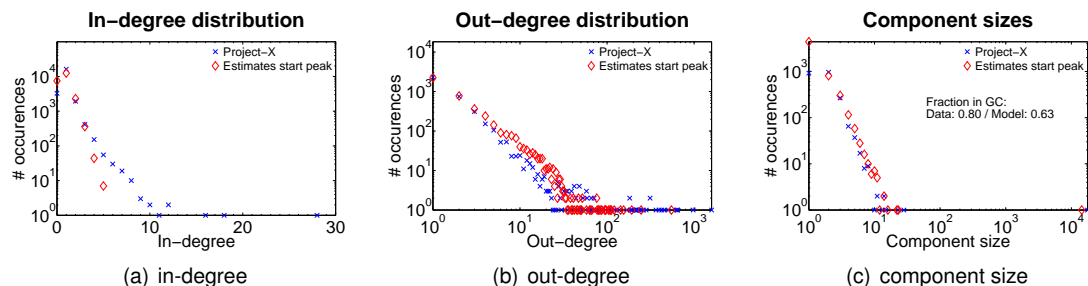


Figure 5.11: Comparison of *Project-X* and simulation using estimates from 2012-9-21 7:00

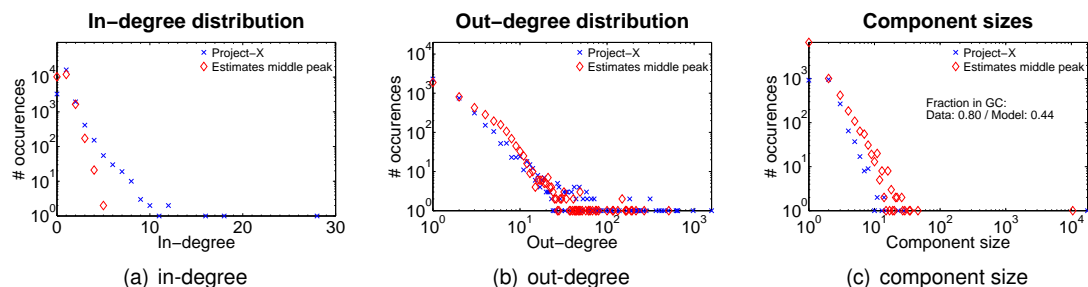


Figure 5.12: Comparison of *Project-X* and simulation using estimates from 2012-9-21 19:00

5.4.2 Turkish-Kurdish dataset

In Figure 5.13 we display the parameter estimates using the tweets available. Again, at the time that the 'other GC' appears, we see a significant change in the estimate of λ . However, this time, also p changes significantly. When we consider the trends in the estimates of λ and p , we find that if λ increases, p decreases. We also find that during every peak in the dataset, the estimate of λ and p change. However, the estimate of q has a very stable progression throughout time. Again, like in Figure 5.7, the estimates stabilize after the last peak.

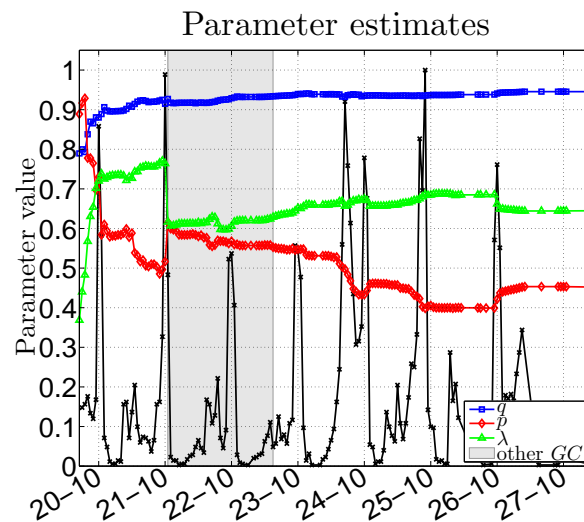


Figure 5.13: Progression of the parameter estimates using the tweets that are available until the indicated time (standard deviations are indicated in red).

The results of the simulations are displayed in Figure 5.14. At the start of the progression, the size of the giant component is highly overestimated. The average degree however, is highly underestimated. Furthermore, the predicted measures display several jumps in their progression. We found that these jumps occur during the peak activity in the dataset. Since this dataset has many peaks in its progression, it is not surprising to see that the predictions vary over time.

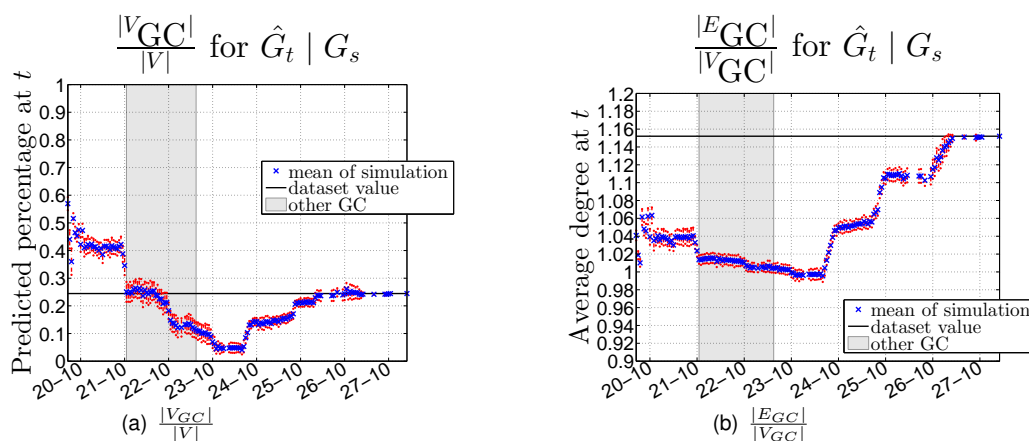


Figure 5.14: Turkish-Kurdish simulations using data that is available to that time.

Then, we compare the first run at the time of the densification (2011-10-20 1:00) with the actual dataset. This is displayed in Figure 5.15. Like in the simulations of the Project-X dataset, the in-degree is underestimated. The out-degree distribution looks more alike, with the exception that the outliers in

the dataset do not appear in the simulations. We also find that the giant component in our simulation is much larger than the actual giant component.

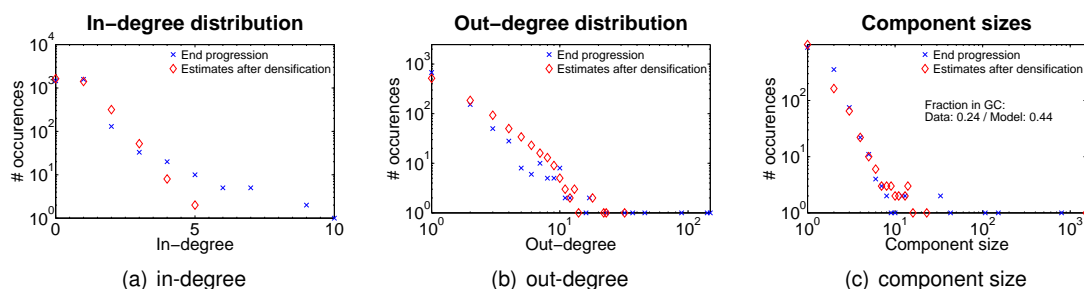


Figure 5.15: Comparison of the *Turkish-Kurdish* dataset and simulation using estimates from 2011-10-20 1:00

5.4.3 WC speedskating dataset

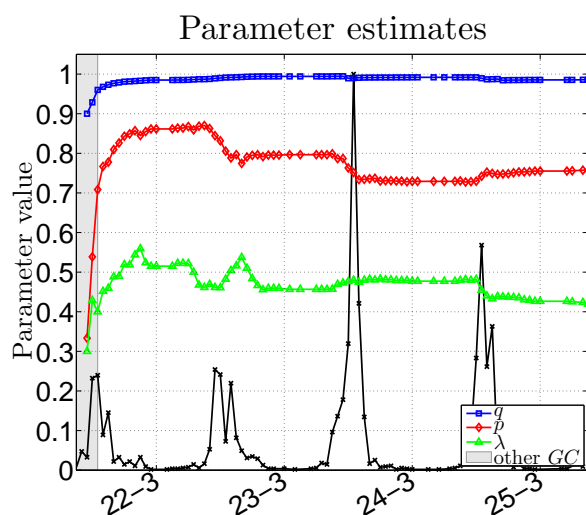


Figure 5.16: Progression of the parameter estimates using the tweets that are available until the indicated time (standard deviations are indicated in red).

In Figure 5.16 we display the parameter estimates using the tweets available. For q , all estimates are very close to 1, which indicates a very star-like structure in the graph, recall that this is the case for this dataset (see Figure 2.10). In this dataset we also see that if λ increases, p decreases. There is however one remarkable fact in the progression of λ . Just after the second peak in the dataset, the value of λ increases shortly, but directly thereafter it decreases again. This indicates that during the second peak of tweets first some components were merged, which was followed by a lot of separate messages.

The results of the simulations are displayed in Figure 5.17. Overall we see that the size of the giant component is underestimated greatly, until the dataset follows the progression of the actual tweets. However, the prediction of the average degree is very good. Since this average degree is close to 1 and we underestimate it slightly, the fact that we underestimate the component size is expected.

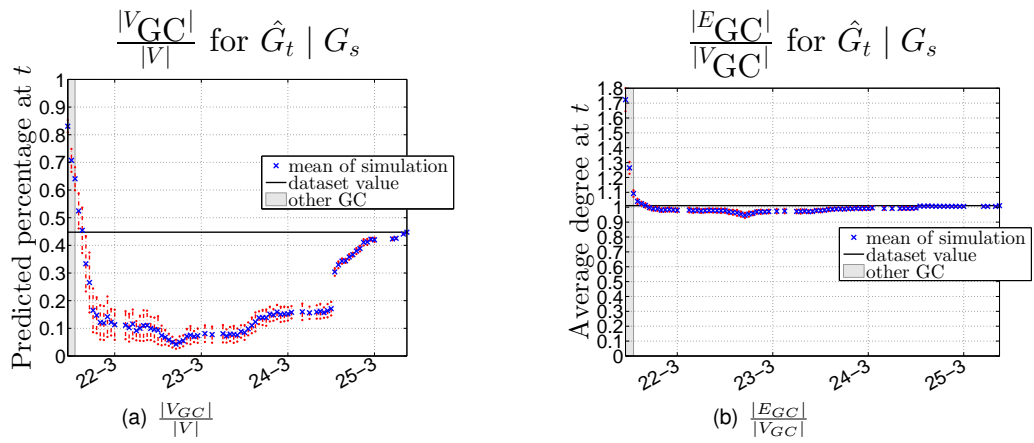


Figure 5.17: WC speedskating simulations using data that is available to that time

CHAPTER 6

CONCLUSIONS AND DISCUSSION

In this chapter we formulate the conclusions of our study and discuss aspects that can be investigated further.

6.1 Conclusions

In this thesis, we aimed to derive predictions for trending behaviour in *Twitter*. We obtained several datasets and used them to investigate which characteristics could be used to model the progression of a topic. Using these characteristics, we formulated a model that can simulate a final progression of a topic using three parameters, being λ , p and q . We derived expressions, which could be used to estimate the parameters using tweet data. Then, we used this model and the available datasets to see how our model performed in comparison to the actual datasets. We find that our model gives reasonable predictions with respect to the average degree in the giant component, which we defined as the largest component in the graph. However, with respect to the size of the giant component, the results vary among the datasets we used.

Since we found that a trend is a progression in which multiple communities are linked together, both the average degree of the giant component and the size of the giant component are properties that our model must be able to predict. Furthermore, we found that the parameters λ and p are the most influential for the end progression of the graph. Since p has a direct influence on the size of the giant component, this is the most decisive parameter with respect to the prediction of trends.

There is one aspect to the prediction of trends that our model does not capture, namely the size of the end progression of a topic. With the addition of this fact, our model can be used for the prediction of trends. Therefore, we conclude that if we develop a way to estimate the size of a retweet graph after a trend, we answered all our research questions in Chapter 1.

6.2 Discussion

In this section we discuss directions for further research in order to improve the current model, which we have not yet included due to the given timeframe. We have considered three datasets in this study. To gain a better insight in the model it is wise to validate it using more datasets. Also, a suggestion for further research into the progression of a single retweet tree is made.

6.2.1 Improvements to the model

In the plots, where we compare the performance of the model with the actual progressions, we see that the in-degree distribution does not match. A way to improve this is to use a preferential attachment mechanism when a target node is chosen. We have tried to include a mechanism based on the in-degree in our model, but it lowered the eccentricity scores considerably. Other options have not been explored so far.

In our work, we use five properties of the graph to compare the real graph with the simulated graph. However, from the assumption that trends emerge as the graph gets denser over time, another measure that could be considered is the clustering coefficient of a node in the network and how this evolves over time. Using this measure will provide a more detailed insight in how the graph becomes denser. In the current version, we cannot distinguish local and global density, using this measure adds to this insight.

The next addition to the model is a time aspect. In our model, we only look at additions in the graph without measuring the time between these additions. We suggest three different ways to extend the model, so that it incorporates a time aspect. The first way to implement this, is to estimate λ_m and λ_r directly (see Table 3.3 for their definitions). This provides more insight in the inter-arrival times of the messages and retweets, which can be used to add a time-aspect to the current model, by assuming both the retweet and the message processes are Poisson processes. The second option is to add the notion of novelty, like Gómez et al. in [20]. We suggest two possibilities to incorporate this. First, in addition to the fact that P_{T_i} is chosen using the size of the message tree T_i , one could introduce the novelty of the message tree at time t $n_{T_i,t}$ that indicates the last time an edge was added to the message tree T_i (denoted by t_{T_i}),

$$n_{T_i,t} = \tau^{t-t_{T_i}}, \quad \tau \in [0, 1].$$

The probability that a tree is chosen can then depend on a weighted selection between novelty and size

$$p_{T_i} = w_1 \cdot \frac{n_{T_i,t}}{\sum_{T_j \in G_t} n_{T_j,t}} + w_2 \cdot \frac{|T_i|}{\sum_{T_j \in G_t} |T_j|}.$$

Second, the probability that a tree is chosen can depend on a combination of novelty and size

$$p_{T_i} = \frac{n_{T_i,t}}{\sum_{T_j \in G_t} n_{T_j,t}} \cdot \frac{|T_i|}{\sum_{T_j \in G_t} |T_j|}.$$

The third aspect is to change the parameters to be time-varying. As can be seen in Figures 5.7, 5.13 and 5.16, the parameters values change over time. Capturing this effect in the model, could improve its output.

In this thesis we have not been able to derive exact equations for the size of the giant component for large values of t . A way to derive such an expression could be to write the model as a two-phase branching process. In this setting the first phase captures the merging of components and the second phase is the branching process that models a single retweet tree. This second phase could be based on the branching process mentioned in [8].

6.2.2 ‘Delayed’ superstar model

In the data we saw that some of the trees, the original message is not the root of the superstar, but this role is fulfilled by a retweeter of that original message, see Figure 6.1(b). We call this a ‘delayed’ superstar progression. In Figure 6.1 we display an example of both types of progressions. The superstar node is indicated by a red circle around the node. The first node of the progression is indicated by a white “1” in the node. These differences arise because we consider directed graphs, instead of undirected graphs as in [8].

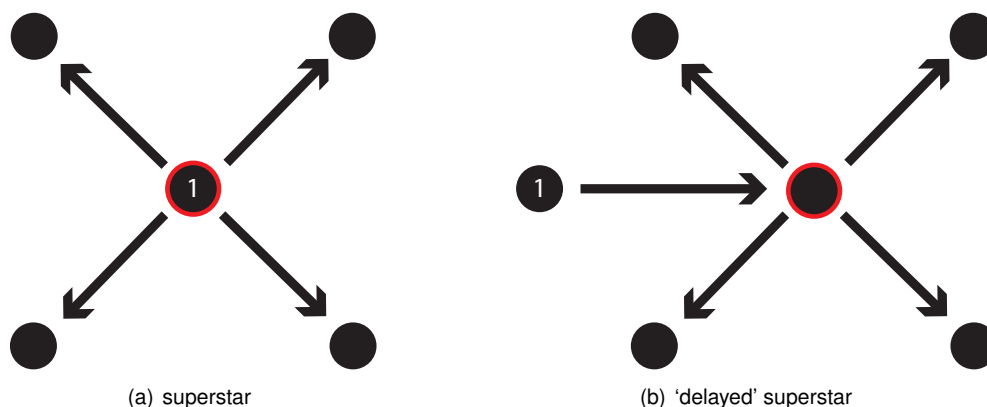


Figure 6.1: Examples of superstar and ‘delayed’ superstar progressions.

BIBLIOGRAPHY

- [1] S. Adali, R. Escriva, M. K. Goldberg, M. Hayvanovych, M. Magdon-Ismael, B. K. Szymanski, W. A. Wallace, and G. Williams. Measuring behavioral trust in social networks. In *Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on*, pages 150–152. IEEE, 2010.
- [2] J. Allan, V. Lavrenko, and H. Jin. First story detection in tdt is hard. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 374–381. ACM, 2000.
- [3] Y. Altshuler, W. Pan, and A. Pentland. Trends prediction using social diffusion models. *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 97–104, 2012.
- [4] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576. Association for Computational Linguistics, 2011.
- [5] S. Ardon, A. Bagchi, A. Mahanti, A. Ruhela, A. Seth, R. M. Tripathy, and S. Triukose. Spatio-temporal analysis of topic popularity in twitter. *arXiv preprint arXiv:1111.2904*, 2011.
- [6] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- [7] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. In *Proceedings of the Fifth International AAI Conference on Weblogs and Social Media (ICWSM11)*, 2011.
- [8] S. Bhamidi, J. M. Steele, and T. Zaman. Twitter event networks and the superstar model. *arXiv preprint arXiv:1211.3090*, 2012.
- [9] D. Bhattacharya and S. Ram. Sharing news articles using 140 characters: A diffusion analysis on twitter. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 966–971. IEEE, 2012.
- [10] B. Bollobás, O. Riordan, J. Spencer, G. Tusnády, et al. The degree sequence of a scale-free random graph process. *Random Structures & Algorithms*, 18(3):279–290, 2001.
- [11] H. Bouma, O. Rajadell, D. Worm, C. Versloot, and H. Wedemeijer. On the early detection of threats in the real world based on open-source information on the internet. In *International Conference on Information Technologies and Security (ITSEC)*, 2012.
- [12] C. Budak, D. Agrawal, and A. El Abbadi. Structural trend analysis for online social networks. *Proceedings of the VLDB Endowment*, 4(10):646–656, 2011.
- [13] A. E. Cano, S. Mazumdar, and F. Ciravegna. Social influence analysis in microblogging platforms—a topic-sensitive based approach. *Semantic Web Interoperability, Usability, Applicability*, 2013.
- [14] C. Castillo, M. Mendoza, and B. Pobleto. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.
- [15] F. Chung, S. Handjani, and D. Jungreis. Generalizations of polya’s urn problem. *Annals of combinatorics*, 7(2):141–153, 2003.
- [16] P. Cogan, M. Andrews, M. Bradonjic, G. Tucci, W. S. Kennedy, and A. Sala. Reconstruction and analysis of twitter conversation graphs. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, pages 25–31. ACM, 2012.

- [17] M. De Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher. How does the data sampling strategy impact the discovery of information diffusion in social media. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 34–41, 2010.
- [18] S. Doan, B.-K. H. Vo, and N. Collier. An analysis of twitter messages in the 2011 tohoku earthquake. *arXiv preprint arXiv:1109.1618*, 2011.
- [19] M. Ester, A. Frommelt, H.-P. Kriegel, and J. Sander. Algorithms for characterization and trend detection in spatial databases. In *Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining, New York, NY*, pages 44–50, 1998.
- [20] V. Gómez, H. J. Kappen, N. Litvak, and A. Kaltenbrunner. A likelihood-based framework for the analysis of discussion threads. *World Wide Web*, pages 1–31, 2012.
- [21] P. Hall. On the rate of convergence of moments in the central limit theorem for lattice distributions. *Transactions of the American Mathematical Society*, 278(1):169–181, 1983.
- [22] N. A. Heard, D. J. Weston, K. Platanioti, and D. J. Hand. Bayesian anomaly detection methods for social networks. *The Annals of Applied Statistics*, 4(2):645–662, 2010.
- [23] T.-A. Hoang and E.-P. Lim. Virality and susceptibility in information diffusions. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.
- [24] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, pages 57–58. ACM, 2011.
- [25] E. Hsu, A. Minich, and D. Ong. Cascade analysis with limited network data. http://snap.stanford.edu/class/cs224w-2011/proj/daniel89_Finalwriteup_v2.pdf.
- [26] B. Huberman, D. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *Available at SSRN 1313405*, 2008.
- [27] C. Hui, Y. Tyshchuk, W. A. Wallace, M. Magdon-Ismail, and M. Goldberg. Information cascades in social media in response to a crisis: a preliminary model and a case study. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 653–656. ACM, 2012.
- [28] B. Klein, X. Laiseca, D. Casado-Mansilla, D. López-de Ipiña, and A. Nespral. Detection and extracting of emergency knowledge from twitter streams. *Ubiquitous Computing and Ambient Intelligence*, pages 462–469, 2012.
- [29] K. Kuldeep, C. Natarajan, and R. Karthik. Tweetstrap: Apriori prediction of retweet counts. <http://blogs.ischool.berkeley.edu/i290-abdt-s12/projects/predicting-tweet-popularity/>.
- [30] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev. Prediction of retweet cascade size over time. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2335–2338. ACM, 2012.
- [31] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.
- [32] J. Lanagan and A. F. Smeaton. Using twitter to detect and tag important events in live sports. In *Proceedings of AAAI*, 2011.
- [33] T. Lane and C. E. Brodley. Temporal sequence learning and data reduction for anomaly detection. *ACM Transactions on Information and System Security (TISSEC)*, 2(3):295–331, 1999.
- [34] R. Lee and K. Sumiya. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, pages 1–10. ACM, 2010.

- [35] W. Lee and D. Xiang. Information-theoretic measures for anomaly detection. In *Security and Privacy, 2001. S&P 2001. Proceedings. 2001 IEEE Symposium on*, pages 130–143. IEEE, 2001.
- [36] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web*, pages 251–260. ACM, 2012.
- [37] K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, 2010.
- [38] J. Li, W. K. Cheung, J. Liu, and C. Li. On discovering community trends in social networks. In *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on*, volume 1, pages 230–237. IET, 2009.
- [39] P. Melville, K. Subbian, C. Perlich, R. Lawrence, and E. Meliksetian. A predictive perspective on measures of influence in networks. In *Proceedings of the Workshop on Information in Networks*, 2010.
- [40] M. Motoyama, B. Meeder, K. Levchenko, G. M. Voelker, and S. Savage. Measuring online service availability using twitter. In *Proceedings of the 3rd conference on Online social networks*, pages 13–13. USENIX Association, 2010.
- [41] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, 2011.
- [42] A.-M. Popescu and M. Pennacchiotti. Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1873–1876. ACM, 2010.
- [43] G. Rattananaritnont, M. Toyoda, and M. Kitsuregawa. A study on characteristics of topicspecific information cascade in twitter. In *Forum on Data Engineering (DE2011)*, 2011.
- [44] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704. ACM, 2011.
- [45] E. Sadikov and M. M. M. Martinez. Information propagation on twitter. *CS322 Project Report*, 2009.
- [46] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.
- [47] A. Signorini, A. M. Segre, and P. M. Polgreen. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PLoS One*, 6(5):e19467, 2011.
- [48] K. Subbian and P. Melville. Supervised rank aggregation for predicting influence in networks. *arXiv preprint arXiv:1108.4801*, 2011.
- [49] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *Data Mining, Fifth IEEE International Conference on*, pages 8–pp. IEEE, 2005.
- [50] T. Takahashi, S. Abe, and N. Igata. Can twitter be an alternative of real-world sensors? *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments*, pages 240–249, 2011.
- [51] J. Weng and B.-S. Lee. Event detection in twitter. *Proc. of ICWSM*, 2011.
- [52] S. Wu, C. Tan, J. Kleinberg, and M. Macy. Does bad news go away faster. In *In Proceedings of the International Conference on Weblogs and Social (ICWSM)*, 2011.

- [53] Z. Xu and Q. Yang. Analyzing user retweet behavior on twitter. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 46–50. IEEE, 2012.
- [54] Y. Yamaguchi, T. Takahashi, T. Amagasa, and H. Kitagawa. Turank: Twitter user ranking based on user-tweet graph analysis. *Web Information Systems Engineering–WISE 2010*, pages 240–253, 2010.
- [55] M.-C. Yang, J.-T. Lee, S.-W. Lee, and H.-C. Rim. Finding interesting posts in twitter based on retweet graph analysis. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1073–1074. ACM, 2012.
- [56] H. Zhang, Q. Zhao, H. Liu, K. xiao, J. He, X. Du, and H. Chen. Predicting retweet behavior in weibo social network. *Web Information Systems Engineering-WISE 2012*, pages 737–743, 2012.
- [57] Z. Zhou, R. Bandari, J. Kong, H. Qian, and V. Roychowdhury. Information resonance on twitter: watching iran. In *Proceedings of the First Workshop on Social Media Analytics*, pages 123–131. ACM, 2010.
- [58] A. Zubiaga, D. Spina, V. Fresno, and R. Martínez. Classifying trending topics: a typology of conversation triggers on twitter. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2461–2464. ACM, 2011.