

Ain Shams University  
Faculty of Computer and Information Sciences  
Computer Science Department



---

## **Developing a Predictive Model for Message Propagation on Online Social Networks**

---

A thesis submitted to the department of computer science, faculty  
of computer and information sciences, Ain Shams university, in  
partial fulfilment of the requirements for the degree of Doctor of  
Philosophy in computer and information sciences

**By:**

**Sarah Abdelwahab Ali Elsharkawy**

M.Sc. in Computer Science,  
Faculty of Computer and Information Sciences,  
Ain Shams University.  
Cairo, Egypt

**Supervised By:**

**Prof. Dr. Mohamed Ismail Roushdy**

Head of Computer Science Department and Former Dean,  
Faculty of Computer and Information Sciences,  
Ain Shams University

**Dr. Ghada Nasr Ali Hassan**

Associate Professor in Computer Science Department,  
Faculty of Computer and Information Sciences,  
Ain Shams University

**Dr. Tarek Mohamed Nabhan**

Research Director,  
Research and Development Department,  
ITWORX

Cairo - 2018



# Abstract

In online social networks such as Twitter, tweeting allows users to share a variety of content to their own followers. As tweets are retweeted from user to user, large cascades of tweets propagation are formed. The growth of cascades over time signals the popularity or lack thereof of the subject matter. The k-core of an information graph is a common measure of a node connectedness in diverse applications. The k-core decomposition algorithm categorizes nodes into k-shells based on their connectivity. Previous research claimed that the super-spreaders are those located at the k-core of a social graph and the nodes become of less importance as they get assigned to a k-shell away from the k-core.

A meme represents an idea or a topic that spreads among users of an online social network. Current research on modelling information diffusion in social media focuses on studying retweet cascades of individual tweets independently. However, as a meme spreads, it evolves, and users adopt the meme in varying manners. While retweet cascades can model the propagation of a single piece of information among users, they are not useful in studying the propagation of the whole meme.

In this thesis, we aim to study the information diffusion from a wider perspective where the information propagation of a meme is tracked rather than individual tweets. And also, investigate the influence effect of the super-spreaders, located at the k-core, on the meme cascade growth.

First, the cascade growth of retweet cascades and the various features that govern the growth are studied. We pose the question of whether the same feature set can be used for cascade growth prediction of any dataset on Twitter. Two types of growth prediction are addressed: structural and temporal. First, a definition of structural and temporal growth is devised. Then, an approach to select the best of these features based on the dataset for better accuracy results is proposed. We present and discuss the results of the most discriminating features

---

in predicting cascades’ growth and provide evidence that the pre-selection of features improved the accuracy of the prediction task on the datasets. Moreover, an evidence that the features governing the cascade growth vary from one dataset to another is found.

Next, we generalize the modelling of retweet cascades to a modelling of the diffusion of a meme. To construct the meme adoption graph (MAG), messages related to a meme are identified from the social network stream. Then, a recent clustering algorithm is utilized to automatically extract and cluster tweets. Next, three epidemic cascade construction models are evaluated and compared to construct the MAG and represent a meme diffusion. Also, a set of structural characteristics derived from the MAG that describe the underlying meme adoption pattern are proposed. An empirical study, using four real-world Twitter datasets, is performed to demonstrate the effectiveness of the proposed MAG.

Moreover, we work towards evaluating the influence span of the social media super-spreaders, located at the  $k$ -core, in terms of the number of  $k$ -shells that their influence is capable of reaching. Our methodology is based on the observation that the  $k$ -core size is directly correlated to the graph size under certain conditions. These conditions are explained and the correlation is utilized to assess the effectiveness of the  $k$ -core nodes for influence dissemination. The results of the carried out experiments show a high correlation between the  $k$ -core size and the sizes of the inner  $k$ -shells in the examined datasets. However, the correlation starts to decrease in the outer  $k$ -shells. Further investigations have shown that the  $k$ -shells, that were less correlated, exhibited a higher presence of spam accounts.

Finally, the effectiveness of using the  $k$ -core nodes, as seed nodes, for influence maximisation is inspected. A measure is proposed to estimate the relative strength of the  $k$ -core as an influence source among other sources of influence contributing to the cascade development. And, we propose combining that measure along with the correlation between the inner  $k$ -core size and the cascade size to determine the influence domination of the  $k$ -core nodes, and hence the effectiveness of targeting these specific nodes for influence maximization.

---

# Publications

- Sarah Elsharkawy, Ghada Hassan, Tarek Nabhan, Mohamed Roushdy, “Studying the K-core Influence Dissemination in Twitter Cascades”, The International Conference on Artificial Intelligence Applications and Innovations (AIAI). Rhodes, Greece. 2018.
- Sarah Elsharkawy, Ghada Hassan, Tarek Nabhan, Mohamed Roushdy. (2017) “Effectiveness of the K-core Nodes as Seeds for Influence Maximisation in Dynamic Cascades”. International Journal of Computers, 2, 187-194.
- Sarah Elsharkawy, Ghada Hassan, Tarek Nabhan, Mohamed Roushdy, “On the Reliability of Cascade Size as a Virality Measure”, Proceedings of the European Conference on Electrical Engineering and Computer Science. Bern, Switzerland. 2017.
- Sarah Elsharkawy, Ghada Hassan, Tarek Nabhan, Mohamed Roushdy, “Towards Feature Selection for Cascade Growth Prediction on Twitter”, Proceedings of the 10th International Conference on Informatics and Systems (INFOS). Cairo, Egypt. 2016.
- Sarah Elsharkawy, Ghada Hassan, Tarek Nabhan, Mohamed Roushdy, “Modelling Meme Adoption Pattern on Online Social Networks”, International Journal of Web Intelligence, 2018. **(Pending)**

---

---

# Acknowledgements

*My deep gratitude, appreciation and sincerest thanks go to Prof.Dr.Mohamed Roushdy, former dean of the faculty of computer and information sciences, Ain Shams University, for his guidance, assistance and advice throughout the thesis.*

*I would like to express my sincere gratitude to my advisor Dr.Ghada Hassan for the continuous support of my Ph.D study and related research, for her patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and of writing this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.*

*I also would like to express my deepest appreciation and sincerest thanks to Dr.Tarek Nabhan for proposing the idea of this thesis and providing continuous constructive feedback from an industrial perspective. He is always committed to provide his guidance in a very short time. His fruitful discussions guided me from the first day in this thesis.*

*Besides my advisors, I would like to thank Prof.Dr.Abdelbadie Salem who provided me with an opportunity to publish a research paper in an international scientific journal.*

*Last but not the least, I would like to thank my family: my husband, my two beloved daughters and my parents for supporting me spiritually throughout writing this thesis and my life in general.*

---

---



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Overview . . . . .	1
1.2	Problem Definition and Motivation . . . . .	2
1.3	Contributions . . . . .	4
1.4	Graphs and Notations . . . . .	6
1.5	Thesis Outline . . . . .	11
<b>2</b>	<b>Review of Literature</b>	<b>13</b>
2.1	Information Diffusion and Prediction of Message Propagation . . .	13
2.2	Influential Spreaders and Influence Maximisation . . . . .	15
2.3	Cascade Modelling and Meme Identification . . . . .	16
2.4	Multiple Sources of Influence and Spam . . . . .	19
<b>3</b>	<b>Data Collection and Dataset Description</b>	<b>21</b>
3.1	Twitter API methods . . . . .	23
3.1.1	Collecting Tweets . . . . .	23
3.1.2	Collecting the Retweeters of a Tweet . . . . .	25
3.1.3	Collecting a User's Information . . . . .	25
3.1.4	Collecting the Followers of a User . . . . .	26
3.2	Further details . . . . .	27
3.3	Twitter Crawling for Dataset Collection . . . . .	28
3.4	Datasets Description . . . . .	28
3.4.1	Collected Tweets . . . . .	29
3.4.2	Collected Retweets . . . . .	29
3.4.3	Collected Users . . . . .	29
3.4.4	Collected User Followers . . . . .	36
3.5	Challenges in the Dataset Collection . . . . .	37
<b>4</b>	<b>Cascade Growth Prediction</b>	<b>39</b>
4.1	Problem Definition . . . . .	40

---

## CONTENTS

---

4.2	Cascade Growth Definition . . . . .	41
4.2.1	Cascade Growth Features . . . . .	42
4.2.2	Feature Selection Approach . . . . .	44
4.3	Experiments . . . . .	46
4.4	Results . . . . .	47
4.4.1	Size Growth Prediction . . . . .	48
4.4.2	Temporal Growth Prediction . . . . .	49
4.5	Discussion and Conclusion . . . . .	51
<b>5</b>	<b>Meme Cascade Modelling</b>	<b>53</b>
5.1	Problem Formulation . . . . .	55
5.1.1	Memeprints: Identifying a Meme . . . . .	55
5.1.2	Modelling Meme Adoption . . . . .	57
5.2	Meme Adoption Graph (MAG) . . . . .	59
5.2.1	Susceptible-Infected ( $MAG_{SI}$ ) . . . . .	59
5.2.2	Susceptible-Infected-Recovered ( $MAG_{SIR}$ ) . . . . .	60
5.2.3	Susceptible-Infected-Susceptible ( $MAG_{SIS}$ ) . . . . .	60
5.2.4	Evaluation of the three models . . . . .	62
5.3	Characteristics of Meme Adoption Pattern . . . . .	62
5.3.1	Comprehensive Properties of MAGs . . . . .	62
5.3.2	Distributional properties of MAG nodes . . . . .	64
5.4	Case Study and Results . . . . .	65
5.4.1	RC Model Vs. MAG . . . . .	65
5.4.2	Meme Adoption Pattern: Structural Properties of MAGs . . . . .	66
5.4.3	Meme Adoption Pattern: Distributional properties of meme adopters . . . . .	69
5.4.4	Meme Adoption Pattern: Tracking of MAGs . . . . .	70
5.5	Summary and Conclusion . . . . .	76
<b>6</b>	<b>The K-core Influence Dissemination</b>	<b>77</b>
6.1	Relationship Between K-core Size and Graph Size . . . . .	79
6.1.1	Conditions for the Presence of Correlation . . . . .	80
6.1.2	Correlation on Synthetic Graphs . . . . .	80
6.1.3	Focusing on Synthetic Scale-Free Power-Law Degree Graphs . . . . .	86
6.1.4	DD Similarity in Real-life Dynamic Cascades . . . . .	87
6.2	Experiments and Results . . . . .	89
6.2.1	Correlation on Twitter Datasets . . . . .	89
6.2.2	The Spam Effect on the Correlation . . . . .	90
6.2.3	Simulation of Crowd-Turfing . . . . .	92
6.3	Discussion and Conclusion . . . . .	95

---

<b>7</b>	<b>Influence Maximisation in Dynamic Cascades</b>	<b>97</b>
7.1	The Effectiveness of the $k_d$ -core as an Influence Source . . . . .	98
7.2	Experiments and Results . . . . .	100
7.2.1	Evaluating the $k_d$ -core as an Influence Source . . . . .	101
7.2.2	The Spam Effect as an External Source of Influence . . . .	102
7.3	Discussion and Conclusion . . . . .	103
<b>8</b>	<b>Summary and Conclusions</b>	<b>105</b>
8.1	Summary . . . . .	105
8.2	Conclusions . . . . .	106
8.2.1	Predicting the Retweet Cascade Growth . . . . .	106
8.2.2	Modelling the Meme Cascade . . . . .	107
8.2.3	Evaluating the Influence Effect of the K-core users on the Meme Cascade Growth . . . . .	107
8.3	Future Work . . . . .	108
8.3.1	Enriching MAGs with Sentiment . . . . .	108
8.3.2	Prediction of MAG Growth . . . . .	108
8.3.3	Weighting Influence Effectiveness of Nodes . . . . .	109
8.3.4	Early Prediction of Spam . . . . .	109
<b>A</b>	<b>Python Coding for Graphs</b>	<b>127</b>
A.1	Gephi: Graph Visualization Software . . . . .	127
A.2	NetworkX . . . . .	128
A.2.1	Reading Graphs . . . . .	128
A.2.2	Writing Graphs . . . . .	129
A.2.3	Creating Graphs . . . . .	129
A.2.4	Graph Properties . . . . .	130
A.2.5	Random Graphs . . . . .	131
A.3	Correlation Between K-core Size and K-shell Size in Python . . . .	133

## CONTENTS

---

---

# List of Tables

3.1	Crawling Details . . . . .	29
3.2	Samples of Tweets from Tsunami . . . . .	30
3.3	Samples of Tweets from Royal Baby . . . . .	31
3.4	Samples of Tweets from Tamarod . . . . .	32
3.5	Samples of Tweets from Tagarod . . . . .	33
3.6	Samples of Retweets . . . . .	34
3.7	Samples of user profiles . . . . .	35
3.8	Samples of user followers . . . . .	37
4.1	List of Content features . . . . .	42
4.2	List of author features . . . . .	43
4.3	List of retweeters features . . . . .	44
4.4	List of structural features . . . . .	44
4.5	List of temporal features . . . . .	45
4.6	Structural Growth Prediction Results of <i>Tamarod</i> . . . . .	49
4.7	Structural Growth Prediction Results of <i>Tagarod</i> . . . . .	49
4.8	Temporal Growth Prediction Results of <i>Tamarod</i> . . . . .	51
4.9	Temporal Growth Prediction Results of <i>Tagarod</i> . . . . .	51
5.1	Graphical properties of RC and MAG on <i>Tagarod</i> dataset . . . . .	67
5.2	Structural properties of MAGs on the four datasets using SI model of epidemics. . . . .	68
5.3	Snapshots of the $MAG_{SI}$ , $MAG_{SIR}$ , and $MAG_{SIS}$ on the first, fourth and eighth day of the <i>Tagarod</i> dataset respectively. . . . .	71
6.1	Graphs of different degree distributions and their k-cores . . . . .	82
6.2	Graphs of different degree distributions and their k-cores . . . . .	83
6.3	Graphs of different degree distributions and their k-cores . . . . .	84
6.4	Graphs of different degree distributions and their k-cores . . . . .	85
6.5	Synthetic graphs grouped based on power-law exponent of node DD. . . . .	87

6.6	Mean and standard deviation of the exponent for the fitted power-law curve of the node degrees of the snapshots taken for each dataset.	90
6.7	Spearman's correlation coefficients ( $r_s$ ) between $k_d$ -core size and graph size of each dataset. . . . .	90
6.8	Spearman's correlation coefficients between $k_d$ -core size and each outer k-shell size measured on all snapshots of each dataset. . . .	91
6.9	Spearman's Correlation Coefficients between $k_d$ -core size and the $MAG_{SIS}$ size. The number of fake tweets injected is a percentage value of the cascade size of the first snapshot in each dataset . . .	94
7.1	Spearman's correlation coefficients ( $r_s$ ) between the $k_d$ -core size and the cascade size, and the average percentage ratio ( $S$ ) of $k_d$ -core successors of each dataset. . . . .	101
7.2	The percentage of spam accounts and their successors present in the datasets. . . . .	102

# List of Figures

1.1	A graph consisting of 6 nodes and 7 edges. . . . .	6
1.2	Clustering Coefficient (C) of Purple Node . . . . .	7
1.3	(a) A binomial degree distribution of a network with 10,000 nodes and average degree of 10. (b) A power law degree distribution of a network with 10,000 nodes and average degree of around 7. . . . .	8
1.4	k-core Decomposition Analysis . . . . .	9
1.5	The yellow node has a high betweenness centrality, and acts as a bridge between subgroups in the network. This figure is taken from <a href="http://www.knicreative.com/tag/degree-centrality/">http://www.knicreative.com/tag/degree-centrality/</a> . . . . .	9
1.6	Graphical representation of a network with five communities and a high modularity. . . . .	11
5.1	Architecture of the tweets clustering approach. . . . .	56
5.2	Retweet cascades for tweets $T_{u_0}(m_i)$ , $T_{u_1}(m_i)$ , $\dots$ , $T_{u_5}(m_i)$ . . . . .	58
5.3	<i>Tagarod</i> dataset modelled at the end of the meme tracking period. The $MAG_{SI}$ model on the left vs. the RC model on the right. . . . .	67
5.4	Probability distribution of the difference between out-degree and in-degree of nodes of the four datasets. . . . .	72
5.5	Probability distribution of the clustering coefficients of the four datasets. . . . .	73
5.6	Probability distribution of the normalized betweenness centrality of the four datasets. . . . .	74
5.7	Monitoring the evolution of the number of nodes in MAGs. A comparison between the SI, SIR, and SIS models. . . . .	74
5.8	Monitoring the evolution of the average node degree in MAGs. A comparison between the SI, SIR, and SIS models. . . . .	75
5.9	Monitoring the evolution of the graph diameter in MAGs. A comparison between the SI, SIR, and SIS models. . . . .	75