

# ANALYSIS OF THE SPREAD OF INFORMATION ON TWITTER AND ITS APPLICATION TO INTERNET CONTENT DISTRIBUTION

AMIT RUHELA



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
INDIAN INSTITUTE OF TECHNOLOGY DELHI  
MAY 2016

© Indian Institute of Technology Delhi (IITD), New Delhi, 2016.  
All rights reserved.

# ANALYSIS OF THE SPREAD OF INFORMATION ON TWITTER AND ITS APPLICATION TO INTERNET CONTENT DISTRIBUTION

by

**AMIT RUHELA**

Department of Computer Science and Engineering

Submitted

in fulfillment of the requirements of the degree of

**Doctor of Philosophy**

to the



**Indian Institute of Technology Delhi**

**May 2016**

To my family for their love, support and patience

# Certificate

This is to certify that the thesis titled “**ANALYSIS OF THE SPREAD OF INFORMATION ON TWITTER AND ITS APPLICATION TO INTERNET CONTENT DISTRIBUTION**” being submitted by Mr. **Amit Ruhela** to the **Indian Institute of Technology Delhi**, for the award of the degree of **Doctor of Philosophy**, is a record of bona-fide research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted to any other university or institute for the award of any degree or diploma.

Dr. Aaditeshwar Seth  
Assistant Professor  
Department of Computer Science and Engineering  
Indian Institute of Technology Delhi, New Delhi 110 016

Date :  
New Delhi

# Acknowledgments

I would like to take this opportunity to thank everyone who contributed to the success of my PhD dissertation. First, I would like to thank my supervisor Prof. Aaditeshwar Seth, who with his insightful suggestions, technical guidance and constructive feedback made my PhD experiences a unique one. He gave me complete freedom in my research work. Moreover, he always supported me in the technical and administrative matters that considerably simplified my PhD journey.

I express my gratitude for Prof. Pankaj Jalote who encouraged me to do a PhD and supported me in the initial years. I am also thankful to Dr. Anirban Mahanti and Prof. Amitabha Baghchi who helped me throughout my research work in IIT Delhi by providing day-to-day suggestions, feedbacks, and encouragement. I would also like to thank Prof. Vinay Joseph Riberio, Prof. S. Arun Kumar and Dr. Amit A. Nanavati, who gave useful feedbacks in my SRC presentations.

I would also like to thank Prof. Naveen Garg, Prof. Preeti Ranjan Panda and Prof. Subodh Kumar, who were PhD coordinators at various points of time. The PhD seminar series initiated by them in our department was extremely helpful in getting diversified outlooks on my work.

I am grateful to my seniors Neeraj Goyal, Anant Vishnoi, Vaibhav Jain and Vikram Goyal for their encouragement and guidance during PhD. I would like to express my gratitude to Dr. Pankaj Dalela and Dr. Ashu Verma with whom I had many fruitful informal discussions on various matters related to PhD. They cannot be thanked enough for the emotional support during the troughs of my PhD life.

Discussions with my PhD colleagues Rudra Mohan Tripathi, Chinmay Narayan, Shibashis Guha, Yamuna Prasad, Namita Sharma, Tarun Beri and Dipanjan Chakraborty

helped me at various points. Long discussions with Rudra Mohan Tripathi about different research possibilities in Online Social Networks laid a foundation in my research aptitude. Along with the colleagues, I am also thankful to numerous unknown reviewers of my conference papers for providing their feedback about my research ideas.

I would like to thank members of the lab staff Vandana Ahulwalia, S S Negi, M. Rathinam, Rajesh Kumar, Arun Kumar and Suresh Kumar, who made available their support in number of ways. I am also thankful to office staff and CSE department for helping me in various ways.

I express my deepest regards to Shri Vipin Tyagi, Shri Praveen Mathur, Dr. Anand Srivastav, Ravi Gupta and Yogesh Kumar who are from C-DOT where I work full time. They have always provided me the necessary support to pursue PhD alongside the official responsibilities. I am also thankful to my colleagues Priyatosh Mandal, Abhay Gupta, Sumit Saharana, Ajay Kumar, Chetan Selwal, Pramod Jhaldiyal and Abhimanyu Popli for supporting me both technically as well as morally throughout my PhD journey.

I dedicate this thesis to my father Late Sh. Sumer Singh Ruhela whose support and encouragement has always been an inspiration to me. Even during last days of his life, my father was extremely caring and enthusiastic for my research work. The endless love, immense patience and moral support of my mother (Smt. Krishna Ruhela), wife (Pooja Ruhela) and daughter (Dhani) throughout my PhD life cannot be expressed in words. I am thankful to my sisters and their in-laws who have supported me a lot in fulfilling various social obligations.

Last but not the least, I owe my deepest gratitude to the Almighty, who makes everything happen, who brings the ideas in one's mind, creates an environment to cherish those ideas, and motivates one to implement the ideas.

# Abstract

With rapidly growing traffic volumes on the Internet, service providers are finding it increasingly hard to manage delivery of content to end users. We observe that a large portion of the traffic is contributed and accessed by Online Social Networking (OSN) websites, and we attempt to leverage this fact to build better content placement and caching strategies for Content Delivery Networks (CDNs). Through a large measurement study of OSN datasets, we show that highly popular topics cross regional geographic boundaries, and time-zone differences can be used to potentially improve content pre-fetching policies. Through an event-based analysis, we show that topic popularity peaks when multiple small components in the social graph coalesce together to form a single giant component. We also show that to track topic trends, it is sufficient to track a small percentile of the most popular users rather than track all users. Finally, we apply these insights to design OSN-aided caching policies in CDNs, but we discover that simple LRU caching does quite well unless cache sizes are very small. This is a significant negative result that can help guide other researchers interested in applying OSN signals to improve Internet performance. Our contributions also go beyond just scoping the potential of using OSN signals to aid content distribution – our insights can find applications in the placement of advertisements on OSN websites, information recommendation, and other areas of work we hope to explore in the future.

# Contents

|                                                                                                         |             |
|---------------------------------------------------------------------------------------------------------|-------------|
| <b>List of Figures</b>                                                                                  | <b>xvii</b> |
| <b>List of Tables</b>                                                                                   | <b>xix</b>  |
| <b>1 Introduction</b>                                                                                   | <b>1</b>    |
| 1.1 Organization of the thesis . . . . .                                                                | 4           |
| <b>2 Related Work</b>                                                                                   | <b>5</b>    |
| 2.1 Related work on spatio-temporal analysis of topic popularity: . . . . .                             | 5           |
| 2.2 Related work on finding events from tweets: . . . . .                                               | 7           |
| 2.3 Related work on the characteristics of Twitter users: . . . . .                                     | 8           |
| 2.4 Related work on usage of OSN signals in CDN: . . . . .                                              | 8           |
| <b>3 Spatio-Events based analysis of Topic popularity in Twitter</b>                                    | <b>11</b>   |
| 3.1 Introduction . . . . .                                                                              | 11          |
| 3.2 Methodology . . . . .                                                                               | 13          |
| 3.2.1 Available datasets . . . . .                                                                      | 13          |
| 3.2.2 Limitations of datasets . . . . .                                                                 | 14          |
| 3.2.3 Data extraction . . . . .                                                                         | 14          |
| 3.3 Events analysis . . . . .                                                                           | 21          |
| 3.4 Geographical analysis . . . . .                                                                     | 25          |
| 3.5 Concluding remarks . . . . .                                                                        | 29          |
| <b>4 The rich and middle classes on Twitter: Are popular users indeed different from regular users?</b> | <b>31</b>   |
| 4.1 Introduction . . . . .                                                                              | 32          |
| 4.2 Datasets and Definitions . . . . .                                                                  | 33          |
| 4.2.1 Defining the popularity of users . . . . .                                                        | 36          |
| 4.2.2 Defining events and event phases . . . . .                                                        | 37          |
| 4.3 Results and Discussion . . . . .                                                                    | 39          |
| 4.3.1 Volume of tweeting . . . . .                                                                      | 39          |
| 4.3.2 Early adopters . . . . .                                                                          | 41          |
| 4.3.3 Engagement with the event . . . . .                                                               | 41          |
| 4.3.4 Participation across different event phases . . . . .                                             | 43          |
| 4.3.5 Influence on growth rates . . . . .                                                               | 45          |

|          |                                                                                              |           |
|----------|----------------------------------------------------------------------------------------------|-----------|
| 4.3.6    | Content copying characteristics . . . . .                                                    | 47        |
| 4.3.7    | Time-delay characteristics of content copying . . . . .                                      | 49        |
| 4.4      | Discussion and conclusion . . . . .                                                          | 51        |
| <b>5</b> | <b>Towards the use of Online Social Networks for efficient Internet content distribution</b> | <b>55</b> |
| 5.1      | Spatial predictor . . . . .                                                                  | 55        |
| 5.2      | Temporal predictors . . . . .                                                                | 56        |
| 5.3      | Social network predictor . . . . .                                                           | 59        |
| 5.4      | Summary . . . . .                                                                            | 61        |
| <b>6</b> | <b>The Scope for Online Social Network aided caching in web CDNs</b>                         | <b>63</b> |
| 6.1      | Introduction . . . . .                                                                       | 63        |
| 6.2      | Methodology . . . . .                                                                        | 65        |
| 6.2.1    | Datasets . . . . .                                                                           | 66        |
| 6.2.2    | Surrogate server positioning . . . . .                                                       | 66        |
| 6.2.3    | Cache sizes . . . . .                                                                        | 66        |
| 6.2.4    | Workload generation . . . . .                                                                | 67        |
| 6.2.5    | Events based caching algorithm . . . . .                                                     | 68        |
| 6.2.6    | System architecture for events based simulations . . . . .                                   | 69        |
| 6.3      | Simulation results . . . . .                                                                 | 70        |
| 6.3.1    | Influence of popular topics . . . . .                                                        | 70        |
| 6.3.2    | Other workloads . . . . .                                                                    | 75        |
| 6.4      | Discussion . . . . .                                                                         | 79        |
| 6.5      | Concluding remarks . . . . .                                                                 | 80        |
| <b>7</b> | <b>Discussions and open problems</b>                                                         | <b>81</b> |
| 7.1      | Conclusions . . . . .                                                                        | 81        |
| 7.2      | Contributions of the thesis . . . . .                                                        | 81        |
| 7.3      | Open problems . . . . .                                                                      | 83        |
|          | <b>Appendices</b>                                                                            | <b>85</b> |
|          | <b>Appendix A Events detection architecture</b>                                              | <b>87</b> |
| A.1      | Event detection algorithm . . . . .                                                          | 88        |
| A.1.1    | Based on inter-arrival time of tweets: . . . . .                                             | 88        |
| A.1.2    | Based on frequency of tweets: . . . . .                                                      | 92        |
| A.2      | Detection of event phases . . . . .                                                          | 93        |
| A.2.1    | Three-phase partitioning of events : . . . . .                                               | 93        |
| A.2.2    | Five-phase partitioning of events : . . . . .                                                | 94        |
|          | <b>Appendix B Reciprocity</b>                                                                | <b>97</b> |
|          | <b>Bibliography</b>                                                                          | <b>99</b> |

# List of Figures

|      |                                                                                                                                                                                                                                 |    |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1  | Overview of thesis works . . . . .                                                                                                                                                                                              | 2  |
| 3.1  | Comparison of OpenCalais input length . . . . .                                                                                                                                                                                 | 18 |
| 3.2  | Frequency distribution of topics . . . . .                                                                                                                                                                                      | 20 |
| 3.3  | Event phases: IAT and User counts . . . . .                                                                                                                                                                                     | 22 |
| 3.4  | Aggregate analysis of topic spread using events based approach: Single giant component of users who talk on a topic forms as the topics go in peak phase of the event . . . . .                                                 | 23 |
| 3.5  | Aggregate analysis of topic spread using events based approach: The decreased value of conductance in peak phase of the events show that disjoint clusters of users merge to form giant component . . . . .                     | 24 |
| 3.6  | Aggregate analysis of topic spread using events based approach: Popular topics cross large number of regional boundaries. . . . .                                                                                               | 24 |
| 3.7  | Aggregate analysis of topic spread using events based approach: Popular topics are nourished by users that have large out-degree. . . . .                                                                                       | 25 |
| 3.8  | Spatial analysis of topic spread: Popular topics cross large number of regional boundaries . . . . .                                                                                                                            | 26 |
| 3.9  | Popularity vs edges crossing for sampled topics . . . . .                                                                                                                                                                       | 27 |
| 3.10 | Spatial analysis of topic spread : In aggregate analysis with 40 topics, the fraction of edges crossing regional boundaries is larger for popular topics than the medium and less popular topics. . . . .                       | 28 |
| 3.11 | Spatial analysis of topic initiation: Higher number of topics are initiated from the countries that have large number of users. . . . .                                                                                         | 28 |
| 3.12 | Spatial analysis of topic initiation: Countries which have large number of heavily followed users initiate large number of topics. . . . .                                                                                      | 29 |
| 4.1  | Cumulative distribution of Tweets wrt followers count . . . . .                                                                                                                                                                 | 37 |
| 4.2  | An example of event detection for topic “AbkiBaarModiSarkar” . . . . .                                                                                                                                                          | 38 |
| 4.3  | CDF plot of rank scores on the basis of tweet volume in the politics dataset. . . . .                                                                                                                                           | 40 |
| 4.4  | (a) Relative time at which popular and medium users start tweeting on topics from the start of the events (b) Relative time at which popular and ordinary users start tweeting on topics from the start of the events . . . . . | 42 |
| 4.5  | CDF plot of time spent by popular and ordinary users in political events . . . . .                                                                                                                                              | 43 |
| 4.6  | Participation of users in subsequent phases of the political events . . . . .                                                                                                                                                   | 44 |

|      |                                                                                                                                                      |    |
|------|------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 4.7  | Temporal plot of events on the basis of ‘Growth Rate’ and ‘Participation of Popular users’ for the politics dataset . . . . .                        | 46 |
| 4.8  | Production of tweets in the political dataset by very popular users, medium popular users, ordinary users and inactive users . . . . .               | 47 |
| 4.9  | Time delay(in minutes) for the 50th percentile of retweets between very popular, medium popular, ordinary and inactive users in the politics dataset | 49 |
| 4.10 | Participation of popular users in the growth phase of events Vs. the popularity of events . . . . .                                                  | 52 |
| 5.1  | Geographical dispersion of topic “Michael Jackson” with time . . . . .                                                                               | 56 |
| 5.2  | Temporal pattern of topic “MichaelJackson” in USA and India . . . . .                                                                                | 57 |
| 5.3  | Periodic topics . . . . .                                                                                                                            | 59 |
| 5.4  | Ephemeral vs stable topics . . . . .                                                                                                                 | 59 |
| 5.5  | Slow vs sharp growing popularity . . . . .                                                                                                           | 60 |
| 5.6  | Social cohesion strength of UGC producers . . . . .                                                                                                  | 60 |
| 6.1  | Time-Delay Distribution of Retweets . . . . .                                                                                                        | 69 |
| 6.2  | Architecture for Events based Simulations . . . . .                                                                                                  | 70 |
| 6.3  | Rank-Frequency plot of Consumption Workload . . . . .                                                                                                | 71 |
| 6.4  | Exponential within 2 Days: 15 Servers . . . . .                                                                                                      | 72 |
| 6.5  | Average time difference between requests on same the contents . . . . .                                                                              | 73 |
| 6.6  | Average index difference between requests on the same contents . . . . .                                                                             | 73 |
| 6.7  | Comparison between LRU and Events for the removal of top k Topics . .                                                                                | 74 |
| 6.8  | Each topic is discussed on how many servers . . . . .                                                                                                | 76 |
| 6.9  | Cache performance by changing count of Surrogate Servers : SN based workload with Exponential Consumption in 7 Days . . . . .                        | 77 |
| 6.10 | One Object per topic Vs Multiple Objects per Topic . . . . .                                                                                         | 78 |
| 6.11 | SN vs Non-SN Model . . . . .                                                                                                                         | 79 |
| A.1  | Block Diagram of Events Detection Modules . . . . .                                                                                                  | 88 |
| A.2  | State diagram of the events detection . . . . .                                                                                                      | 89 |
| A.3  | Events Detection on topic “Michael Jackson” . . . . .                                                                                                | 91 |
| A.4  | State Diagram of Events Detection . . . . .                                                                                                          | 92 |
| A.5  | An example on Event Detection for topic “AbkiBaarModiSarkar” . . . . .                                                                               | 94 |
| A.6  | Phases of an Event . . . . .                                                                                                                         | 95 |
| B.1  | Reciprocity . . . . .                                                                                                                                | 98 |
| B.2  | Reverse Reciprocity . . . . .                                                                                                                        | 98 |

# List of Tables

|     |                                                                       |    |
|-----|-----------------------------------------------------------------------|----|
| 3.1 | Summary of available datasets . . . . .                               | 14 |
| 3.2 | Geographical statistics . . . . .                                     | 16 |
| 3.3 | Tweets statistics . . . . .                                           | 17 |
| 3.4 | Top entry of each OpenCalais category . . . . .                       | 19 |
| 4.1 | Datasets attributes (M=Million) . . . . .                             | 34 |
| 6.1 | Time spent in the cache by the topics . . . . .                       | 74 |
| 6.2 | Removal of top K topics in SN based workload with 5 servers . . . . . | 75 |
| 6.3 | Relative Cache Performance . . . . .                                  | 76 |
| A.1 | Parameters used in events detection algorithm . . . . .               | 91 |
| A.2 | Variables of events detection algorithm . . . . .                     | 93 |