

# VU Research Portal

## Social Data Analysis

ten Thij, M.C.

2018

### **document version**

Publisher's PDF, also known as Version of record

[Link to publication in VU Research Portal](#)

### **citation for published version (APA)**

ten Thij, M. C. (2018). *Social Data Analysis: Dynamics of real-time data*. [PhD-Thesis - Research and graduation internal, Vrije Universiteit Amsterdam].

### **General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

### **Take down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

### **E-mail address:**

[vuresearchportal.ub@vu.nl](mailto:vuresearchportal.ub@vu.nl)

# Contents

<b>List of Abbreviations</b>	<b>v</b>
<b>1 Motivation and Outline</b>	<b>1</b>
1.1 Outline . . . . .	4
<b>2 Preliminaries</b>	<b>7</b>
2.1 Stochastic Processes . . . . .	7
2.2 Random Graphs . . . . .	15
2.3 Natural Language Processing . . . . .	17
<b>I Diffusion Analysis of Social Data</b>	<b>21</b>
<b>3 Introduction</b>	<b>23</b>
<b>4 Modelling Information Diffusion in Twitter</b>	<b>27</b>
4.1 Retweet Graph Process . . . . .	28
4.2 Growth of the Retweet Graph . . . . .	32
4.3 Simplified Random Graph Model . . . . .	37
<b>5 Degree Distribution of the Retweet Graph Model</b>	<b>39</b>
5.1 Degree Difference Equation . . . . .	40
5.2 Retweet Graph Model Without Merging . . . . .	44
5.3 Degree Distribution of the Simplified Model . . . . .	49
5.4 Tail Behaviour of the Simplified Model . . . . .	56
5.5 Degree Distribution of the Retweet Graph Model . . . . .	58

## Contents

---

<b>6</b>	<b>Component Sizes in the Random Graph Model</b>	<b>61</b>
6.1	Incremental Growth of the Component Size Distribution . . . . .	62
6.2	Retweet Graph Model Without Merging . . . . .	63
6.3	General Component Size Distribution . . . . .	65
6.4	Size of the Largest Connected Component . . . . .	68
<b>7</b>	<b>Retweet Graph Model in Practice</b>	<b>75</b>
7.1	Real-world Data Sets . . . . .	76
7.2	Parameter Estimation and Simulation . . . . .	81
<b>II</b>	<b>Activity Analysis of Social Data</b>	<b>83</b>
<b>8</b>	<b>Introduction</b>	<b>85</b>
<b>9</b>	<b>Activity Dynamics in Wikipedia</b>	<b>89</b>
9.1	Data Set . . . . .	90
9.2	Page-view Statistics . . . . .	91
9.3	Promoted Articles . . . . .	95
9.4	Circadian Patterns Correction . . . . .	96
9.5	Page-view Behaviour Model . . . . .	99
<b>10</b>	<b>Activity Dynamics in Twitter</b>	<b>115</b>
10.1	Data Set . . . . .	116
10.2	Tweet Activity Statistics . . . . .	120
10.3	Tweet Diffusion Patterns . . . . .	122
<b>III</b>	<b>Real-time Analysis of Social Data</b>	<b>127</b>
<b>11</b>	<b>Introduction</b>	<b>129</b>
<b>12</b>	<b>Horticulture Analytics</b>	<b>133</b>
12.1	Product Sales . . . . .	134
12.2	Event/Campaign Tracking . . . . .	136
<b>13</b>	<b>Hortiradar</b>	<b>139</b>
13.1	Streamer . . . . .	140
13.2	Preprocessor . . . . .	141
13.3	Analyser . . . . .	142
13.4	Visualiser . . . . .	143
13.5	Statistics . . . . .	145
<b>14</b>	<b>Clustering Content in Real-time</b>	<b>147</b>
14.1	Obtaining Long-term Stories . . . . .	148
14.2	The Storify Algorithm . . . . .	149
14.3	Comparing the Storify Algorithm to Post-hoc Clustering . . . . .	152

14.4 Optimising the Storify Algorithm . . . . .	156
<b>Publications</b>	<b>161</b>
<b>Bibliography</b>	<b>163</b>
<b>Online References</b>	<b>179</b>
<b>Summary</b>	<b>181</b>
<b>Samenvatting</b>	<b>185</b>
<b>About the Author</b>	<b>189</b>

## Contents

---