

Contents

1	Introduction	1
1.1	Research Question and Challenges	2
1.2	Objectives	3
1.2.1	Main objective	3
1.2.2	Specific objectives	3
1.3	Contributions	3
1.4	Methodology	4
1.5	Outline of this work	5
2	Background and Related Work	6
2.1	Data Analysis Techniques	6
2.1.1	Vector Space Model	6
2.1.2	Clustering and community detection in graphs	7
2.1.3	Graph Clustering Evaluation Metrics	8
2.1.4	Topic models and event detection	9
2.2	Content diversity and characterization in social networks	10
2.3	Visualizations of News Media Relationships	11
3	Theoretical Framework	12
3.1	Graph model	12
3.2	Similarity measures	13
3.2.1	Vocabulary Similarity	13
3.2.2	Topic Similarity	14
3.2.3	Temporal Correlation	14
3.2.4	Ownership similarity	17
3.2.5	Follower similarity	18
3.3	Community discovery	19
3.3.1	Preprocessing	19
3.3.2	Community analysis	19
4	Experimental Methodology	21
4.1	Methodology overview	21
4.2	Dataset	22
4.2.1	News documents source	22
4.2.2	Outlet information sources	24
4.3	Implementation and experiments	24
4.3.1	Similarity measures and internal analysis	25

4.3.2	Similarity distribution exploration	26
4.3.3	Community discovery	28
4.3.4	Community analysis	28
5	Results and Analysis	32
5.1	Ownership structure	33
5.2	Follower similarity	34
5.3	Vocabulary similarity	38
5.4	Topic similarity	41
5.5	Temporal correlation	44
5.5.1	Penta term set	44
5.5.2	President term set	47
5.6	Comparison	49
5.7	Insights	50
6	Conclusions	51
6.1	Applications	51
6.2	Limitations	54
6.3	Extensions and Improvements	54
	Bibliography	55
	Appendices	61
A	User distribution for each community structure	61
B	Similarity graph visualizations for each explored similarity measure	64
C	Lists of stop words	68
C.1	Stop words included from NLTK	68
C.2	Manually included stop words	70