University of Kentucky
UKnowledge

Theses and Dissertations--Business Administration

Business Administration

2016

SOCIAL MEDIA ANALYTICS – A UNIFYING DEFINITION, COMPREHENSIVE FRAMEWORK, AND ASSESSMENT OF ALGORITHMS FOR IDENTIFYING INFLUENCERS IN SOCIAL MEDIA

Shih-Hui Hsiao
University of Kentucky, suade0904@msn.com
Digital Object Identifier: http://dx.doi.org/10.13023/ETD.2016.326

Right click to open a feedback form in a new tab to let us know how this document benefits you.

Recommended Citation

Hsiao, Shih-Hui, "SOCIAL MEDIA ANALYTICS – A UNIFYING DEFINITION, COMPREHENSIVE FRAMEWORK, AND ASSESSMENT OF ALGORITHMS FOR IDENTIFYING INFLUENCERS IN SOCIAL MEDIA" (2016). *Theses and Dissertations--Business Administration*. 8.
https://uknowledge.uky.edu/busadmin_etds/8

This Doctoral Dissertation is brought to you for free and open access by the Business Administration at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Business Administration by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Shih-Hui Hsiao, Student

Dr. Clyde Holsapple, Major Professor

Dr. Kenneth Troske, Director of Graduate Studies

SOCIAL MEDIA ANALYTICS − A UNIFYING DEFINITION, COMPREHENSIVE FRAMEWORK, AND ASSESSMENT OF ALGORITHMS FOR IDENTIFYING INFLUENCERS IN SOCIAL MEDIA

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Business and Economics
at the University of Kentucky

By

Shih-Hui Hsiao

Lexington, Kentucky

Co-Directors: Dr. Clyde W. Holsapple and Dr. Ram Pakath,

Professors of Decision Science & Information Systems

Lexington, Kentucky

Copyright © Shih-Hui Hsiao 2016

ABSTRACT OF DISSERTATION

SOCIAL MEDIA ANALYTICS − A UNIFYING DEFINITION, COMPREHENSIVE FRAMEWORK, AND ASSESSMENT OF ALGORITHMS FOR IDENTIFYING INFLUENCERS IN SOCIAL MEDIA

Given its relative infancy, there is a dearth of research on a comprehensive view of business social media analytics (SMA). This dissertation first examines current literature related to SMA and develops an integrated, unifying definition of business SMA, providing a nuanced starting point for future business SMA research. This dissertation identifies several benefits of business SMA, and elaborates on some of them, while presenting recent empirical evidence in support of foregoing observations. The dissertation also describes several challenges facing business SMA today, along with supporting evidence from the literature, some of which also offer mitigating solutions in particular contexts.

The second part of this dissertation studies one SMA implication focusing on identifying social influencer. Growing social media usage, accompanied by explosive growth in SMA, has resulted in increasing interest in finding automated ways of discovering influencers in online social interactions. Beginning 2008, many variants of multiple basic approaches have been proposed. Yet, there is no comprehensive study investigating the relative efficacy of these methods in specific settings. This dissertation investigates and reports on the relative performance of multiple methods on Twitter datasets containing between them tens of thousands to hundreds of thousands of tweets. Accordingly, the second part of the dissertation helps further an understanding of business SMA and its many aspects, grounded in recent empirical work, and is a basis for further research and development. This dissertation provides a relatively comprehensive understanding of SMA and the implementation SMA in influencer identification.

KEYWORDS: Social Media Analytics, Social Influencers, Opinion Leader, Opinion Mining, Decision Support Systems

Shih-Hui Hsiao

July 26, 2016

Date

SOCIAL MEDIA ANALYTICS − A UNIFYING DEFINITION, COMPREHENSIVE FRAMEWORK, AND ASSESSMENT OF ALGORITHMS FOR IDENTIFYING INFLUENCERS IN SOCIAL MEDIA

By

Shih-Hui Hsiao

Dr. Clyde Holsapple
Co-Director of Dissertation

Dr. Ram Pakath
Co-Director of Dissertation

Dr. Kenneth R. Troske
Director of Graduate Studies

July 26, 2016

ACKNOWLEDGEMENTS

First of all, I would like express my sincerely appreciation to both of my advisors, Dr. Clyde Holsapple and Dr. Ram Pakath. Since I stated my life in US as a PhD student, both of them provide me great training in different disciplines and strong supports for developing my own research interest. Their mentorship helps me improve my capability in conducting research and requires me strong knowledge for teaching in college. I really appreciate their efforts and considerations to help me start my career as a faculty. More import, I thank them for their guidance, encouragement, and support during the development of this work. I am indebted to my committee members: Dr. Chen Chung and Dr. Goldsmith. They have provided, with kindness, their insight and suggestions, which are precious to my work.

Secondly, I would like to thank Dr. Ajay Mehra for helpful advices and inspiring discussions, allowing me to expand my idea to this dissertation and further research projects. His great knowledge in social network leads me to develop my own interest in social media. I also want to thank Po-Chang Su for helping me to develop the programing code in this dissertation. His background in computer science supplements my knowledge in programing and algorithms. Mrs. Jill Westfall and Dr. Merrie Bergmann also provide great contribution in editing the English writing of this work. This dissertation cannot be done without their great supports.

Finally, I want to thanks to my family in Taiwan. The supports from my parents help me go through all the difficult time, living far from home and working alone in US. I would like to express my eternal gratitude to my parents for their everlasting love and support. Thanks to my brother for taking care of my families in Taiwan, which allows me to focus on my study. I also thank to all my friends in US and Taiwan for their supports and companionship. Special thanks are due to Yen-Yao Wang and Wei-Hao Chang for the numerous conversations when I feel weak and powerless. Most importantly, I want to express my sincere gratitude to my fellows, Jae-Young Oh and Zhiguo Yang. It’s my pleasure to work with both of them during my PhD study.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS III

TABLE OF CONTENTS IV

LIST OF TABLES VI

LIST OF FIGURES VII

CHAPTER 1. INTRODUCTION 1

CHAPTER 2. STATE OF THE ART OF BUSINESS SOCIAL MEDIA ANALYTICS 4
2.1. INTRODUCTION 4
2.2. RESEARCH METHODOLOGY 5
2.3. BUSINESS SMA DEFINITION, BENEFITS, AND CHALLENGES 6
2.3.1. Business SMA Definition 6
2.3.2. Business SMA Benefits 9
2.3.3. Business SMA Challenges 12
2.4. PREVIOUS SMA FRAMEWORKS FROM THE LITERATURE 17
2.4.1. The Mayeh et al. Framework (2012) 18
2.4.2. The Sinha et al. Framework (2012) 19
2.4.3. The Stieglitz & Dang-Xuan (2012) & The Stieglitz et al. Framework (2014) 20
2.4.4. The He et al. Framework (2015) 22
2.5. A FRAMEWORK OF SOCIAL MEDIA ANALYTICS-BASED DECISION MAKING 23
2.5.1. Analysis Goal(s) 23
2.5.2. Social Media: Input Data 24
2.5.3. Intelligence 25
2.5.4. Design and Choice 28
2.6. CONCLUDING REMARK 30

CHAPTER 3. AN ASSESSMENT OF ALGORITHMIC SOCIAL INFLUENCER IDENTIFICATION APPROACHES 31
3.1. INTRODUCTION 31
3.2. SOCIAL MEDIA DEVELOPMENT 32
3.3. SOCIAL MEDIA CATEGORIZATION 33
3.4. SOCIAL INFLUENCERS 38
3.5. INFLUENCER IDENTIFICATION 39
3.6. USE OF SMA IN IDENTIFYING INFLUENCERS 41
3.6.1. PageRank Algorithm 43
3.6.2. HITS Algorithm 44
3.6.3. Clustering Algorithm 45
3.6.4. Regression Analysis 46
3.6.5. Centrality Measurement 46
3.6.6. Tag/Topic/Interest –oriented Algorithms 46
3.7. EXPERIMENT DESIGN 47
3.7.1. Data Collection and Description 47
3.7.2. Evaluation Metrics 50

3.7.3. Experiment Procedure *52*
3.8. DISCUSSION 53
3.8.1. Results: Computation Time *53*
3.8.2. Results: The Quality of Identified Influencers *55*
3.8.3. Ensemble Approaches *118*
3.9. SUMMARY 122

CHAPTER 4. CONCLUSION 125
4.1. CONTRIBUTIONS 125
4.2. LIMITATIONS 126
4.3. FUTURE WORKS 127

APPENDICES 128
APPENDIX I: INFLUENCER IDENTIFICATION APPROACHES 128
APPENDIX II: EVALUATION METRICS REVIEW FROM LITERATURE 144
APPENDIX III: EXPERIMENT RESULTS FROM THE TWITTER MARCH MADNESS DATASETS 146
APPENDIX IV: EXPERIMENT RESULTS FROM THE TWITTER KY DERBY DATASET 173

REFERENCES 200

VITA 210

LIST OF TABLES

TABLE 2.1 CATEGORIZATION OF REVIEWED LITERATURE 6
TABLE 2.2 SOCIAL MEDIA ANALYTICS DEFINITIONS 7
TABLE 2.3 RESULTS OF 3RD ANNUAL EMPLOYEE ENGAGEMENT SURVEY (APCO WORLDWIDE & GAGEN MCDONALD, 2011) 11
TABLE 2.4 CURRENT CHALLENGES OF SOCIAL MEDIA ANALYTICS 13
TABLE 3.1 CATEGORIZATION OF SOCIAL MEDIA (MANGOLD & FAULDS, 2009) 34
TABLE 3.2 DEFINITION AND FEATURES OF DIFFERENT TYPES OF SOCIAL MEDIA 36
TABLE 3.3 COMPUTATION TIME FOR FINDING THE TOP *N*% OF INFLUENCERS: TWITTER ELECTION DATASET 53
TABLE 3.4 COVERAGE, LANGUAGE DIFFUSION, AND AGREEMENT RATES OF INFLUENCERS IN INTERSECTION ENSEMBLES: TWITTER ELECTION DATASET 119
TABLE 3.5 COVERAGE, LANGUAGE DIFFUSION, AND AGREEMENT RATES OF INFLUENCERS IN UNION ENSEMBLES: TWITTER ELECTION DATASET 120

LIST OF FIGURES

FIGURE 2.1 VISUALIZING BUSINESS SOCIAL MEDIA ANALYTICS DEFINITION 9
FIGURE 2.2 THE PROPOSED CONCEPTUAL FRAMEWORK (MAYEH ET AL., 2012) 18
FIGURE 2.3 A CONTEMPORARY MODEL OF SMA FOR BEHAVIOR INFORMATICS, HR AND CUSTOMERS (SINHA ET AL., 2012) 19
FIGURE 2.4 SMA FRAMEWORK IN POLITICAL CONTEXTS (STIEGLITZ & DANG-XUAN, 2012) 20
FIGURE 2.5 SOCIAL MEDIA ANALYTICS FRAMEWORK (STIEGLITZ ET AL., 2014) 21
FIGURE 2.6 A SOCIAL MEDIA COMPETITIVE ANALYTICS FRAMEWORK WITH SENTIMENT BENCHMARKS FOR INDUSTRY-SPECIFIC MARKETING INTELLIGENCE (HE ET AL., 2015) 22
FIGURE 2.7 A FRAMEWORK OF SMA-BASED DECISION MAKING 29
FIGURE 3.1 THE HALF-LIFE OF INFORMATION AND INFORMATION DEPTH OF DIFFERENT SOCIAL MEDIA TYPES (ADAPTED FROM WEINBERG & PEHLIVAN, 2011 & HOFFMAN & FODOR, 2010) 35
FIGURE 3.2 SOCIAL INFLUENCERS IDENTIFICATION APPROACHES AND INPUT DATA TYPES 43
FIGURE 3.3 ILLUSTRATION OF HITS ALGORITHM 45
FIGURE 3.4 DATA WINDOW CONSTRUCTIONS 48
FIGURE 3.5 ILLUSTRATION FOR UNDERSTANDING EVALUATION METRICS 51
FIGURE 3.6 COVERAGE RATE FOR DIFFERENT ALGORITHMS: TWITTER ELECTION DATASET 56
FIGURE 3.7 COVERAGE RATE LIFT RATIO CHARTS FOR DIFFERENT ALGORITHMS: TWITTER ELECTION DATASET 60
FIGURE 3.8 COVERAGE RATE OF TOP *N*% OF INFLUENCERS FOR DIFFERENT ALGORITHMS IN DIFFERENT CATEGORIES: TWITTER ELECTION DATASET 64
FIGURE 3.9 LANGUAGE DIFFUSION RATE FOR DIFFERENT ALGORITHMS: TWITTER ELECTION DATASET 77
FIGURE 3.10 LANGUAGE DIFFUSION RATE LIFT RATIO CHARTS FOR DIFFERENT ALGORITHMS: TWITTER ELECTION DATASET 81
FIGURE 3.11 LANGUAGE DIFFUSION RATE OF TOP *N*% OF INFLUENCERS FOR DIFFERENT ALGORITHMS IN DIFFERENT CATEGORIES: TWITTER ELECTION DATASET 85
FIGURE 3.12 AGREEMENT RATE FOR DIFFERENT ALGORITHMS: TWITTER ELECTION DATASET 98
FIGURE 3.13 AGREEMENT RATE LIFT RATIO CHARTS FOR DIFFERENT ALGORITHMS: TWITTER ELECTION DATASET 102
FIGURE 3.14 AGREEMENT RATE OF TOP *N*% OF INFLUENCERS FOR DIFFERENT ALGORITHMS IN DIFFERENT CATEGORIES: TWITTER ELECTION DATASET 106

Chapter 1. Introduction

In 2014, Ellen DeGeneres used a Samsung Galaxy Note 3 smartphone to take a "selfie" while concurrently hosting the Academy Awards and incorporating close to a dozen of Hollywood's top celebrities in the image. She then posted the selfie on Twitter and captioned it: "If only Bradley's arm was longer. Best photo ever." In a short period of time, the tweet went viral, received more than 1.3 million retweets, disrupted Twitter's service for over 20 minutes (Gerick, 2014; Rodgers & Scobie, 2015; Zhu & Chen, 2015), promoted the smartphone worldwide, and caused countless firms to contemplate this phenomenon (and research associated feedback, discussions, and "memes"). However, while such firms have traditionally used statistical methods (e.g., regressions) to analyze social media data and generate new insights, analytics techniques from the quantitative and computational social sciences (e.g., social network analysis) are also now being applied to analyze human social phenomena that is transmitted via social media (Wang et al., 2007).

The significant effects of big data on social media have precipitated a need for superior tools for analyzing social media content, which one group of scholars has started to define as "social media analytics (SMA)." SMA generally (i) focuses on optimizing capabilities associated with managing complex social media data structures (within social media contexts) and (ii) employs up-to-date information technologies to manage the three Vs of big data: volume, velocity, and variety (to support different tasks in different areas (Russom, 2011)). This dissertation focuses specifically on SMA applications in business domains.

Due to the relative infancy of SMA, prior research has generally not yielded a comprehensive view of SMA in business domain (Zeng et al., 2010). Thus, the first part of this dissertation (i) reviews prior SMA literature to develop a comprehensive understanding of it (e.g., the first study analyzes existing SMA definitions and develops a unique definition tailored to business domains), (ii) documents some of the applications of SMA, (iii) articulates some present-day challenges associated with the application of SMA, (iv) introduces a framework for SMA-based decision-making, and (v) fully describes how to use SMA to support decision-making processes in business domains. The second part investigates a specific implementation of business SMA: Social Influencer Identification, which has been studied in varied contexts for decades. For example, (i) in marketing, influencer research has focused on opinion leaders as marketplace influencers (Feick & Price, 1987) and (ii) in education, research has studied the

impact of social influencer on education environments and associated communities (Dill & Friedman, 1979)). Still, social influencer identification, within social media contexts, remains challenging due to the rapid and ongoing generation of social media data; for this reason, academia has developed multiple approaches for supporting influencer identification in social media contexts.

Because there are a variety of influencer identification mechanisms, it can be difficult to select appropriate ones for identifying influencer within specific social media contexts. For instance, differing types of social media (Kaplan & Haenlein, 2010) have generated new types of data structures that require new representations (e.g., text content, network structure, and user activity) and increased the complexity of business SMA (and thus locating influencer). Furthermore, different types of social media (e.g., blog, microblog, and social networking site (SNS)) may require different types of influencer identification approach. Thus, a significant goal of the second part of this dissertation is to (i) evaluate alternate approaches for identifying influencer (in varied and large social media networks), (ii) examine different types of data structures (associated with different types of social media), and (iii) examine the relative performance (i.e., quality and computation time) of several different mechanisms (proposed in the literature).

This dissertation (i) provides a relatively comprehensive understanding of how SMA is applied in business contexts (to help academics attain a clearer understanding of business SMA), (ii) clarifies the applications of (and challenges with) business SMA, and (iii) suggests new directions for the study of business SMA. Future research might focus on individual components that comprise the proposed decision-support framework contained herein (which supports business SMA development); additionally, the assessment of social influencer identification approaches, in this study, yields a relatively precise direction for future SMA research.

Thus, this research helps (i) practitioners with designing SMA-based decision-support systems (based on the decision-making framework described herein); (ii) improve social influencer identification (based on a thorough assessment of multiple mechanisms); and (iii) assist educators via the provision of a (a) paradigm of business SMA development and (b) investigation of business SMA implementations. The dissertation is organized as follows: Chapter 2 presents related literature on SMA, and associated topics (to support a broad understanding of business SMA development); documents current applications and challenges found in the literature (to further detail business SMA applications); and explores the framework of SMA-

based decision-making processes. Chapter 3 explores (i) social influencer-related literature, (ii) mechanisms supporting influencer identification development, (iii) experimental designs that support the assessment of influencer identification approaches, and (vi) discussion of the results from multiple experiments of influencer identification approaches. Chapter 4 concludes, summarizes the main contribution of this dissertation, and suggests future directions for research. The limitations of this dissertation are also discussed in this chapter.

Chapter 2. State of the Art of Business Social Media Analytics

2.1. Introduction

In recent years, social media analytics (SMA) has become highly significant within the diverse field of analytics (Kurniawati et al., 2013). Broadly speaking, SMA applies appropriate analytics capabilities (e.g., sentiment analysis, trend analysis, topic modeling, social network analysis, and visual analytics) to social media content in order to generate specific types of knowledge. Such content can be generated via a variety of social media (Sinha et al., 2012), including (i) blog (e.g., blogger), (ii) microblog (e.g., Twitter), (iii) social bookmarking site (e.g., Delicious), (iv) social networking site (e.g., LinkedIn), (v) review site (e.g., Yelp), and (vi) multimedia sharing site (e.g., YouTube). Currently, SMA takes the approach of "listening" to available social media content (vs. "asking" for user input) and acting upon it. One of the main reasons for the growing interest in SMA is the depth and reach of social media, which is primarily due to content volumes and diffusion speeds.

Still, SMA can be an extremely difficult task in many settings since user-generated social media content is often ad hoc, free-form, and tends to contain both relevant and irrelevant information (from the perspective of specific analytics goals). Nevertheless, analysis of social media data interests many academics and practitioners who are increasingly investing time, money, and effort to pursue SMA-related capabilities. For example, Sterne (2010) mentions that analysis of relevant social media data can benefit businesses seeking to measure customer feedback (e.g., buzz topic trends, volumes of customer buzz (about products or services), buzz diffusion over time, and resultant impacts on sales). This information can help firms improve their marketing strategies.

Due to the relative infancy of SMA, there is a dearth of research yielding comprehensive views of it within business contexts and defining its key characteristics, benefits, limitations, associated strategies, and challenges (e.g., during deployments of associated solutions). Formal studies are thus required to generate new insights and systematic developments in the field. The first part of this dissertation contributes to this by (i) analyzing existing definitions of SMA, (ii) arriving at a specific business-domain definition, (iii) documenting (with recent empirical evidence) some of the business benefits of SMA, (iv) articulating (with empirical support) some of the challenges encountered when applying SMA in business (and other) domains, and (v) recommending solutions for particular contexts. This dissertation ultimately presents a framework

of SMA-based decision-making process (based on Simon's decision making model) to generate a relatively comprehensive view of how to implement SMA in support of business decision-making.

This chapter is structured as follows. The next section briefly describes the approach (i.e., the literature search, filtration, and categorization) utilized in this present, conceptual study. The following section contains various definitions of SMA (found in the extant literature) and presents a definition that is suitable for business contexts. This is followed by an examination of some business-related benefits of SMA and an articulation of challenges associated with applying SMA in business and other contexts. The final section summarizes this work and offers concluding remarks.

2.2. Research Methodology

The first study of this dissertation used Google Scholar for the literature search (i) using all of the words in "social media analytics" and "social media intelligence" and (ii) focusing on a 10-year timeframe (i.e., 2005 through 2015); this yielded 78,300 papers (with stipulated keywords anywhere in the papers). This study subsequently chose papers with the keywords appearing in the title only (which resulted in 215 hits) and eliminated less-productive hits (e.g., duplicates, lecture notes, isolated abstracts, or topics unrelated to businesses), which resulted in a reduced set of 37 papers.

A significant goal of this research was to review the most related and interesting empirical studies; in order to help build up a sufficient and diverse number of current empirical studies, this study conducted another search (i) using "user-generated content" as the key words and (ii) stipulating that these words could appear anywhere in the paper. This search further targeted specific Information Systems journals (*MIS Quarterly, Information Systems Research, Journal of Management Information Systems, Journal of the Association for Information Systems, Decision Support Systems, International Journal of Electronic Commerce, Electronic Markets, and Electronic Commerce Research and Applications*) and the 2010-2015 period. This yielded 838 papers. This study examined these papers and handpicked 24 SMA-related papers, which adapted SMA approaches in business domain and represented a diverse pool of business SMA applications. This study categorized this set of papers into four classes (as shown in Table 2-1).

Table 2.1 Categorization of Reviewed Literature

Category	Description	Number of Studies	Percentage
Empirical Research	Use existing SMA methodologies on social media data to help answer specific research questions.	39	63.93 %
Algorithm/Methodology Design	Develop analysis procedures for SMA that seeks to improve upon extent methods.	15	24.59 %
Conceptual Framework	Develop a framework to help better understand the SMA phenomenon.	13	21.31 %
Case Study	Describe specific scenarios for SMA applications.	4	6.56 %

Some papers fit into more than one category (e.g., a methodology design paper coupled with empirical research). Table 2.1 indicates that the proportion of studies devoted to SMA applications (e.g., empirical studies, case studies, and algorithm/methodology design) far outweighs more comprehensive, conceptual studies that could yield greater insights into the SMA phenomenon as a whole. Thus, this study makes a contribution to the conceptual niche and addresses the dearth of sufficient research to date.

2.3. Business SMA Definition, Benefits, and Challenges

2.3.1. Business SMA Definition

Table 2.2 lists various definitions of SMA that are identified in the reviewed literature. An examination of these definitions reveals that SMA has been defined in terms of the types of activities pursued during SMA life cycles, which include:

- Pre-analytics processing activities – e.g., searching/scanning/monitoring, finding/identifying, collecting, and filtering social media data;
- Analytics processing activities – e.g., assimilating, summarizing, visualizing, analyzing, mining, and generating insights from the data; and,
- Post-analytics processing activities – e.g., interpreting, reporting, dash boarding/alerting, and otherwise utilizing the results of the analytics endeavor.

These activities are not necessarily conducted independently and linearly; often, analysts must repeatedly cycle through and reiterate prior activities during a life cycle to arrive at useful

analytics outcomes. Furthermore, some have determined that SMA (Grubmüller, Götsch, & Krieger, 2013; Grubmüller, Krieger, & Götsch, 2013; Yang et al., 2011) includes a collection of tools, systems, and/or frameworks to facilitate aforementioned activity types. Some researchers (Yang et al., 2011; Zeng et al., 2010) even regard the development and evaluation of such tools, systems, and frameworks as falling within the purview of an SMA definition. Some definitions elaborate on the nature of what is being analyzed (e.g., data on conversations, engagement, sentiment, influence (Sinha et al., 2012; Yang et al., 2011); posts, comments, conversations (Grubmüller, et al., 2013); or, semi-structured and unstructured data (Kurniawati et al., 2013)). This study concurs that certain SMA applications will require further development and evaluation.

Table 2.2 Social Media Analytics Definitions

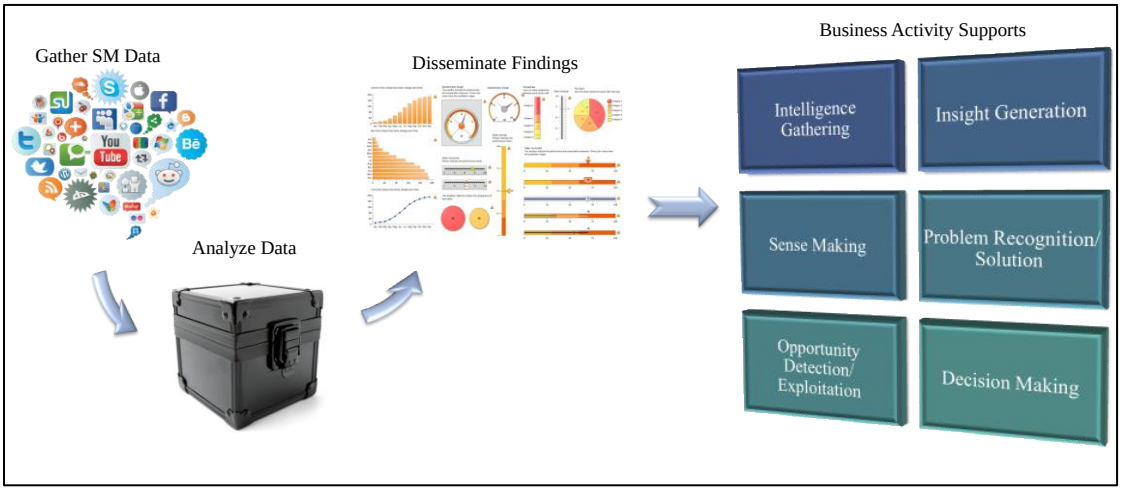
SMA Definition	**Source**
"… developing and evaluating informatics tools and frameworks to collect, monitor, analyze, summarize, and visualize social media data, usually driven by specific requirements from a target application."	(Zeng et al., 2010)
"… developing and evaluating informatics tools and frameworks to measure the activities within social media networks from around the web. Data on conversations, engagement, sentiment, influence, and other specific attributes can then be collected, monitored, analyzed, summarized, and visualized."	(Yang et al., 2011)
"… scanning social media to identify and analyze information about a firm's external environment in order to assimilate and utilize the acquired external intelligence for business purposes."	(Mayeh et al., 2012)
"… measure behavior, conversation, engagement, sentiment, influence, …;" "monitor exchange of information on social networking site."	(Sinha et al., 2012)
"… social listening and measurements … based on user-generated public content (such as posts, comments, conversations in online forums, etc.)" [using SMA tools] "with different features like reporting, dash boarding, visualization, search, event-driven alerting, and text mining."	(Grubmüller, Götsch, et al., 2013)
"Software systems that automatically find filter and analyze user-generated contents produced on social media."	(Grubmüller, Götsch, et al., 2013)
"… the use of analytics-based capabilities to analyze and interpret vast amounts of semi-structured and unstructured data from online sources." "…. provides … insights into …customer values, opinions, sentiments, and perspectives …."	(Kurniawati et al., 2013)

Drawing on Table 2.2, this study advances a business-domain-specific definition of SMA (business SMA) to both facilitate discussions that follow and to provide a useful starting point for those engaged in business SMA research:

"All activities related to Gathering relevant Social Media data, Analyzing the gathered data, and Disseminating findings, to support business activities such as Intelligence Gathering, Problem Recognition, and Opportunity Detection to facilitate Sense Making, Insight Generation, and/or Decision Making in response to sensed business needs."

This business SMA definition not only incorporates the ideas embodied in the prior definitions in Table 2.2, but also lends purpose as to why a business entity might choose to engage in SMA (Figure 2.1). Also inherent in the definition is support for evidence-based problem solving/decision making as advocated by (Grubmüller, Götsch, et al., 2013; Ribarsky et al., 2013). Whereas some authors (e.g., Kurniawati et al., 2013; Mayeh et al., 2012) view business SMA as being synonymous with "customer-centric" SMA, this unifying definition does not preclude the inclusion and analysis of social media data from other business-related entities such as employees, suppliers, retailers, competitors, regulatory bodies, and so forth. This view is shared by Mayeh et al. (2012) with the distinction that they regard a business firm's SMA attempts as being applied only to its external environment. This study, however, contends that a firm's internal environment is also susceptible to SMA, as with monitoring employee/employer-generated internal social media content; the definition accommodates SMA in both the external and internal environs. Finally, the purpose of business SMA is not merely intelligence gathering as Mayeh et al. (2012) contend – it goes beyond intelligence gathering to supporting such activities as insight generation, sense making, problem recognition and solution, opportunity detection and exploitation, and decision making.

Figure 2.1 Visualizing Business Social Media Analytics Definition



2.3.2. Business SMA Benefits

A literature review reveals that business SMA has the potential to provide several benefits to a firm. Specifically, Kurniawati et al. (2013) note the following benefits based on a review of 40 SMA "success stories" (e.g., from IBM, SAS, and SAP): (i) improved marketing strategies (75% (of cases)), (ii) better customer engagement (65%), (iii) better customer service (35%), (iv) better reputation management and brand awareness (30%), (v) product innovation (30%), (vi) business-process improvement (25%), and (vii) discerning new business opportunities (20%). On the other hand, Sinha et al. (2012) discuss the benefits of utilizing behavioral informatics and human resources analytics for recruiting, training, internal communications, employee engagement, talent management, employee/employer branding, and employee life-cycle management. Based on the definition contained herein, this work in this dissertation extends the notions of engagement and service to apply not only to customers but also to employees, business partners, and (in the case of socially conscious firms) societies (i.e., entities affected by such firms). Business SMA (and its benefits) could, in principle, apply to any and all such entities or sectors. Thus, this section briefly describes some benefits along with illustrative, recent empirical research evidence.

1. ***Superior Marketing Strategies*:** Customer-generated content (i) usually contains valuable information about consumer experiences with specific products or services and (ii) is often available on (a) review website (e.g., Epinion.com, Amazon, and Yelp) and (b) personal social networking site (e.g., Facebook).

SMA can provide useful insights for developing and/or refining marketing/sales strategies. Kurniawati et al. (2013) note that most (i.e., 75%) of the reported vendor success stories concern market strategy improvements; furthermore, there is ample empirical work related to studies of this benefit. For example, Hu et al. (2014) evaluate the relative impacts of (i) text sentiments and (ii) star ratings on book sales at Amazon and determine that textual reviews (vs. ratings) directly and significantly impact sales. Interestingly, this is particularly evident with the two most accessible reviews: most helpful and most recent. Additionally, Dellarocas et al. (2010) find that moviegoers show a propensity for reviewing very obscure movies in addition to very popular ones. They hypothesize that user-review volume for lesser-known products may be increased by deliberately obfuscating true volumes of prior movie reviews.

2. ***Better Customer Engagement*:** SMA can be used to identify and target customer values and preferred customer channels for two-way communications and thus enhance Business-to-Customer engagement. (1) Abrahams (2013) evaluates multiple approaches for identifying customer values (e.g., elicitations of customer requirements) associated with automotive components via mining threads in three discussion forums (Honda Tech, Toyota Nation, and Chevrolet Forum). (2) Goh et al. (2013) analyze a clothing retailer's (i) "brand-community" Facebook page (i.e., fan page) and (ii) online community purchase information to show that undirected communication yields superior (i.e., informative and persuasive) C2C communications; however, directed communication yields superior (i.e., more persuasive) marketer-to-consumer (M2C) communication, which is a form of B2C communication.

3. ***Better Customer Service*:** Superior customer service, a goal for many firms in today's hyper-competitive business environments, can potentially be supported by SMA. Hill & Ready-Campbell (2011) describe a genetic algorithm-based opinion mining tool that helps identify effective stock-picking experts from the online Motley Fool CAPS voting site (which contains over 2 million picks of over 770,000 registered users). They show that, as a group, stocks recommended by these experts do better over time (vs. stocks recommended by the S&P 500 or selected via input from all voters on the CAPS site). The net result is (i) stock portfolios that generate superior returns (in this case, for clients of a financial services firm) via opinion mining and (ii) improved customer satisfaction.

4. ***Reputation Management*:** SMA may also be utilized to monitor, maintain, and enhance a firm's reputation (e.g., in association with a brand, product, service, employer, employee, or facility). Deloitte (2012) notes that a growing number of global financial services firms are instituting CRO (chief risk officer) positions and that more CROs are instituting "stress tests" that consider reputational and operational and regulatory risks when assessing a firm's ability to withstand future industry downturns.

   As an example of employer brand-reputation enhancement via the use of social media (APCO Worldwide & Gagen McDonald, 2011), a survey of 1,000 full-time employees (employed for at least one year at firms with at least 500 employees) found that (i) 58% would rather work for a company that uses internal social media (ISM) tools, (ii) 51% said their companies use some form of ISM, (iii) 63% felt that their employers used ISM "well," (iv) 61% felt that collaboration with colleagues was easier with ISM, and (v) 60% felt that use of ISM was indicative of company innovation. Other interesting employer brand reputation-related findings are shown in Table 2.3.

Table 2.3 Results of 3rd Annual Employee Engagement Survey (APCO Worldwide & Gagen McDonald, 2011)

Observation	**Proportion of Respondents At Companies Doing "Well" with ISM**	**Proportion of Respondents At Companies Doing "Fairly or "Poorly" with ISM**
Will likely continue as employee for the foreseeable future	91%	74%
Would likely encourage others to consider employment at company	86%	51%
Would recommend company's products or services	89%	64%
Would give company benefit of the doubt when it's facing litigation/crises	88%	55%
Would purchase company stock	75%	45%

As an instance of reputation management-related research, Amblee & Bui (2011) analyzed the role of electronic word-of-mouth (eWOM) communication in a closed community of low-priced Amazon Shorts e-book readers. They found that eWOM is effective in conveying the reputations of products (i.e., e-books), brands (i.e., authors), and complementary goods (i.e., e-books in the same categories).

5. ***New Business Opportunities:*** There is documented evidence that SMA can also help reveal untapped business opportunities by helping identify new product/service possibilities. For example, Colbaugh & Glass (2011) have developed a method for (i) spotting emerging "memes" (i.e., distinctive phrases that act as "tracers" for topics) and (ii) predicting memes that will propagate wildly and result in significant numbers of discussions of new topics and trends. They argue that such knowledge (e.g., the social network positions of meme originators) is helpful in (i) spotting memes that are likely to propagate wildly, (ii) generating insights, (iii) enhancing B2C communications, and (iv) helping firms determine (a) where to open new retail outlets, (b) what features to include in new products, and (c) which opinion leaders to further engage.

2.3.3. Business SMA Challenges

Like any technology-enabled "big data" solution, business SMA has its own set of challenges; however, as a nascent and developing field, these challenges are also opportunities for further research exploration. In Table 2.4, this chapter summarizes key challenges identified in the literature. For ease of comprehension, this study divides these challenges into two categories: (i) pre-analytics processing activities and (ii) analytics processing activities. The former category includes challenges with specialized context; the processing of free-form, context-sensitive content; data validity concerns pertaining to the use of abbreviations, typos, and questionable credibility; data extraction difficulties (due to data size, data/source variety, and the challenges of separating useful from useless information); and the complexities of processing the nearly continuously streaming flow of data in real time. The latter category focuses on difficulties stemming from the limited life of usable data, its time-varying nature, and methodological issues associated with developing integrative, multidisciplinary, big data-scalable analysis techniques.

In the reviewed literature, there are very few references that specifically identify and address challenges relevant to post-analytics processing. This, of course, could be an artifact of

this particular literature review procedure; however, this study believes that there will very likely be considerable difficulties associated with suitably packaging and disseminating actionable SMA results (with substantial levels of automation and in real-time) due to aforementioned pre-processing and processing challenges. Thus, research on semi-automated and automated, real-time post-SMA processing should experience growing interest in the years ahead by those engaged in business intelligence-related research (of which business SMA is a part).

Table 2.4 Current Challenges of Social Media Analytics

Challenges	**Source**	**Description**
Related to Pre-Analytics Processing Activities		
Context & Structure: free-form statements; unclear broader context	Best et al. (2012)	"…the brevity of most messages, the frequency of data ingest, and the context sensitivity of each message."
	Mosley (2012)	"One major consideration is that social media data tends to be informal…" … the challenge becomes connecting the right set of social media data together to be able to understand the broader context of a conversation."
	Mayeh et al. (2012)	"The unstructured and distributed nature and volume of this information makes the task of extracting useful and practical information challenging."
	Ribarsky et al. (2013)	"Text messages from Twitter, Facebook, and several other social media services have general attributes such as unstructured content…"
Language Use: special symbols; slang use	Zeng et al. (2010)	"Social media applications are a prominent example of human-centered computing with their own unique emphasis on social interactions among users."
	Mosley (2012)	"There are certain symbols that actually do have a meaning and therefore extra care needs to be taken in cleansing the text." "…to understand sentiment would require a more thorough investigation into the ways that users communicate sentiment, and then attempting to capture those sentiments within the data in a structured way."
	Fan & Gordon (2014)	"Language issues add further complications as businesses begin to monitor and analyze social media conversations around the world."

Data Validity: abbreviations, typos, and credibility issues	Asur & Huberman (2010)	"…it was difficult to correctly identify tweets that were relevant to those movies. For instance, for the movie 2012, it was impractical to segregate tweets talking about the movie, from those referring to the year."
	Mayeh et al. (2012)	"Social media data is unstructured, distributed and of uncertain credibility." "Social media data includes spam, which are non-sensible or gibberish text. There are some intentional misspellings used to show commenter's sentiment."
	Mosley (2012)	"…issues with misspellings and abbreviations will be a larger challenge…there are no system edits that ensure the social media data that was captured is accurate, and this may result in false information and statements that are driven by pure emotion rather than fact."
	Ribarsky et al. (2013)	"…intrinsic uncertainty as to the validity of the messages."
Data Extraction: diversity, scope of social media; isolating useful input	Melville et al. (2009)	"An important consideration is to avoid crawling, parsing and storing parts of the blog sub-universe that are irrelevant from a marketing perspective."
	Colbaugh & Glass (2011)	"…most memes receiving relatively little attention and a few attracting considerable interest."
	Chae et al. (2012)	"The relevant messages for situational awareness are usually buried by a majority of irrelevant data."
Streaming Nature: continuously flowing data	Melville et al. (2009)	"However, the set of domains to monitor may change often, requiring classifiers to adapt rapidly with the minimum of supervision."
	Barbieri et al. (2010)	"Stream reasoning moves from this processing model to a continuous model, where tasks are registered and continuously evaluated against flowing data."
	Zeng et al. (2010)	"Social media data are dynamic streams, with their volume rapidly increasing. The dynamic nature of such data and their sheer size pose significant challenges to computing in general and to semantic computing in particular."
	Best et al. (2012)	"The real-time nature of social media analytics for emergency management poses interesting visualization challenges."

Related to Analytics Processing Activities		
Analysis Time Frame: time-varying impacts; limited usable data life	Asur & Huberman (2010)	"For each movie, we define the critical period as the time from the week before it is released, when the promotional campaigns are in full swing, to two weeks after release, when its initial popularity fades and opinions from people have been disseminated."
	Barbieri et al. (2010)	"Data streams are unbounded sequences of time-varying data elements that form a continuous flow of information. Recent updates are more relevant because they describe the current state of a dynamic system."
	Colbaugh & Glass (2011)	"…which will go on to attract substantial attention, and to do so early in the meme lifecycle…although memes typically propagate for weeks, useful predictions can be made within the first twelve hours after a meme is detected."
	Chae et al. (2012)	"There is a need for advanced tools to aid understanding of the extent, severity and consequences of incidents, as well as their time-evolving nature"
	Mosley (2012)	"Trending topics can literally begin in an instant and can become widespread very fast, and if the analysis occurs too long after the topic is trending, it may be too late for the company to do anything useful about it."
	Boden et al. (2013)	"A user can adopt the role of "novice user" the first time she registers with a particular online community, and achieve the role of "expert" months after."
	Ribarsky et al. (2013)	"Events are bursts of activity over a relatively short time period, the time scale depending on the category of the temporal data."
Methodology: integrative, multidisciplinary big data-scalable approaches	Melville et al. (2009)	"Although, clustering and topic modeling techniques can find sets of posts expressing cohesive patterns of discussion, for generating marketing insight we need to identify clusters that are also novel or informative compared to previous streams of discussion."
	Zeng et al. (2010)	"Social media intelligence research calls for highly integrated multidisciplinary research. Although this need has been reiterated often in this growing field, the level of integration in the existing research tends to be low."
	Colbaugh & Glass (2011)	"…in order to identify features of social diffusion which possess predictive power, it is necessary to assess predictability using social and information network models with realistic topologies."
	Boden et al. (2013)	"However, analysts usually face significant problems in scaling existing and novel approaches to match the data volume and size of modern online communities."

Due to the challenges noted in Table 2.4, there have been attempts to articulate partial solutions in particular contexts. This study explores some examples of suggested workarounds associated with some of the challenges suggested by a few authors. For example, with context and structure-related challenges, Mosley (2012) describes a step-by-step process for cleansing and structuring over 68,000 free-form Allstate insurance-related tweets via the use of 116 keywords (associated with the tweets) to examine keyword associations through cluster analysis and association rule mining. However, while the procedure for structuring this type of free-form content (with a maximum of 140 character tweets) is relatively easy, the same cannot be assumed of general, free-form (e.g., blog) content.

Language use is one of the more challenging concerns. Fan & Gordon (2014) allude to the possible use of machine translations to assist with the mining of multilingual content. However, machine translation is still a developing area of research and thus this application is not yet a panacea. Mosley (2012) notes that one solution is to strip gathered tweets of all special symbols (e.g., punctuations, quotation marks, parentheses, and currency symbols); however, (i) such symbols can sometimes convey useful meanings (e.g., the utilizations of smileys) and (ii) significant care should thus be taken when deciding on the symbols that should be retained (vs. stripped away) based on context (e.g., @ and #, in the case of Twitter).

White et al. (2012) have explored data validity and (i) noted that as much as 50% of Twitter content is estimated to be spam (although there is a declining trend) and (ii) observed a large number of identical Tweets from different accounts that were (a) not re-tweets and (b) dispatched within a few hours of one another. An examination of the associated websites led them to conclude that these were filler messages intended to fool the anti-spam defenses of Twitter. Whereas it was possible to detect invalid content in this special circumstance, the circumvention of data validity-related challenges continues to be amongst the most difficult of challenges facing SMA.

The varieties of social media networks (and their copious content) result in unique data extraction challenges. For example, Melville (2009) has contemplated how to avoid crawling, parsing, and storing irrelevant parts of a blog sub-universe and recommended a “focused snowball sampling” procedure with a (i) text classifier (to determine relevant links in a given blog) and (ii) web crawler (to add the blogs associated with these links). This process is reiterated with each new identified blog until some predetermined “degrees of separation” count has been attained.

Best (2012) describes features of a prototype system (the Scalable Reasoning System (SRS)) for real-time visualizations of emergency management system (EMS)-related Twitter information (for use by the City of Seattle). They noted a greater importance of real-time (vs. continual) updates despite the fact that tweets are often continuously flowing on such occasions. Their solution was to embed a clock in the user interface to remind users to request timely, periodic refreshes. A refresh indicator also displays the number of new tweets accumulated in the interim (i.e., since the last refresh). Users can ask for refreshes (based on elapsed time, number of new tweets, or both).

Ribarsky et al. (2013) explore challenges introduced by the Analysis Time Frame and the impacts of time-varying behaviors of tweet metadata (e.g., retweet count, follower count, and favorite count) on an SMA attempt. All metadata values are only accurate as of the moment of tweet collection; thus, a tweet collected shortly after generation may show a retweet count value of zero or a follower count of 10. However, if the same tweet were captured later, these values could be substantially higher. Furthermore, if one were to repeatedly capture the same tweets at various epochs (to circumvent aforementioned difficulties), one could end up with a vastly larger dataset and the tweet-gathering process could require many months. The authors thus devise a mitigating solution for long-running tweet-collection settings that takes advantage of the fact that an original tweet is embedded within retweets of the same tweet. One could focus on gathering only (i) popular and influential tweets that tend to get retweeted repeatedly and (ii) retweets capturing time-varying metadata.

As Table 2.4 notes, there is often a need for utilizing integrative, multidisciplinary, big data-scalable methods during SMA exercises. Zeng et al. (2010) note that although this is now a recognized fact, substantial progress has been lacking in recent years. Still, Yang et al. (2011) attained successful integration via mining web forums maintained by hate groups opposing radical opinions. They utilized both machine learning and a semantic-oriented approach to extract four types of features characterizing radical opinions in such forums. They next applied three classification methods (i.e., Support Vector Machine, Naive Bayes, and AdaBoost) to classify new posts as being radical or benign.

2.4. Previous SMA Frameworks from the Literature

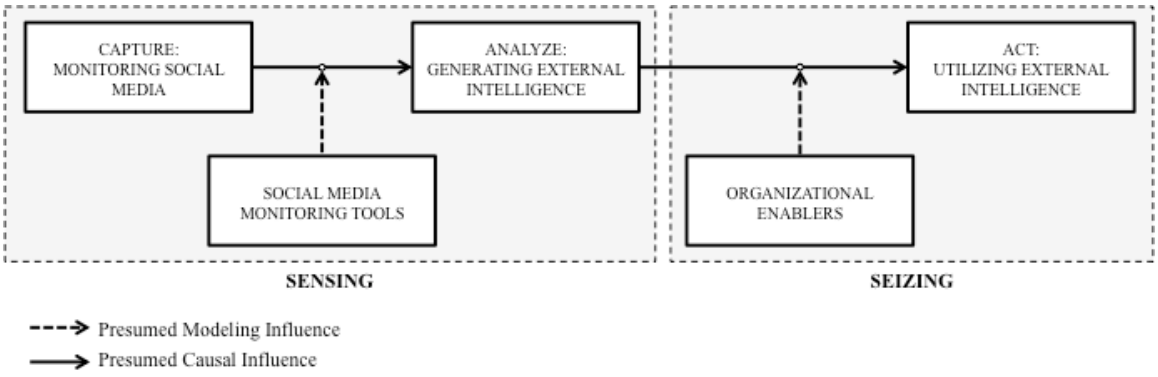
In this section, the study first reviews five social media analytics frameworks propsed in the last few years. This study then proposes a framework based on Simon's model of decision-

making to articulate SMA processes as they pertain to business setting. This study unifies previous framework into a relatively general and comprehensive framework, complementing previous frameworks with precise steps and data analytics capabilities to support business decision-making.

2.4.1. The Mayeh et al. Framework (2012)

Mayeh et al. (2012) have studied the potential utility of social media data for firms gathering external intelligence (e.g., from customers, competitors, suppliers, partners, industries, and technologies). The authors adapt the concept of dynamic capability and design a conceptual framework (Figure 2.2) based on the concept of dynamic capabilities (Teece et al., 1997), which is "the capacity of an organization to purposefully create, extend, or modify its resource base" (Helfat et al., 2009). Teece (2007) defines dynamic capabilities as encompassing opportunity sensing, opportunity seizing, and threats management/transformation.

Figure 2.2 The Proposed Conceptual Framework (Mayeh et al., 2012)

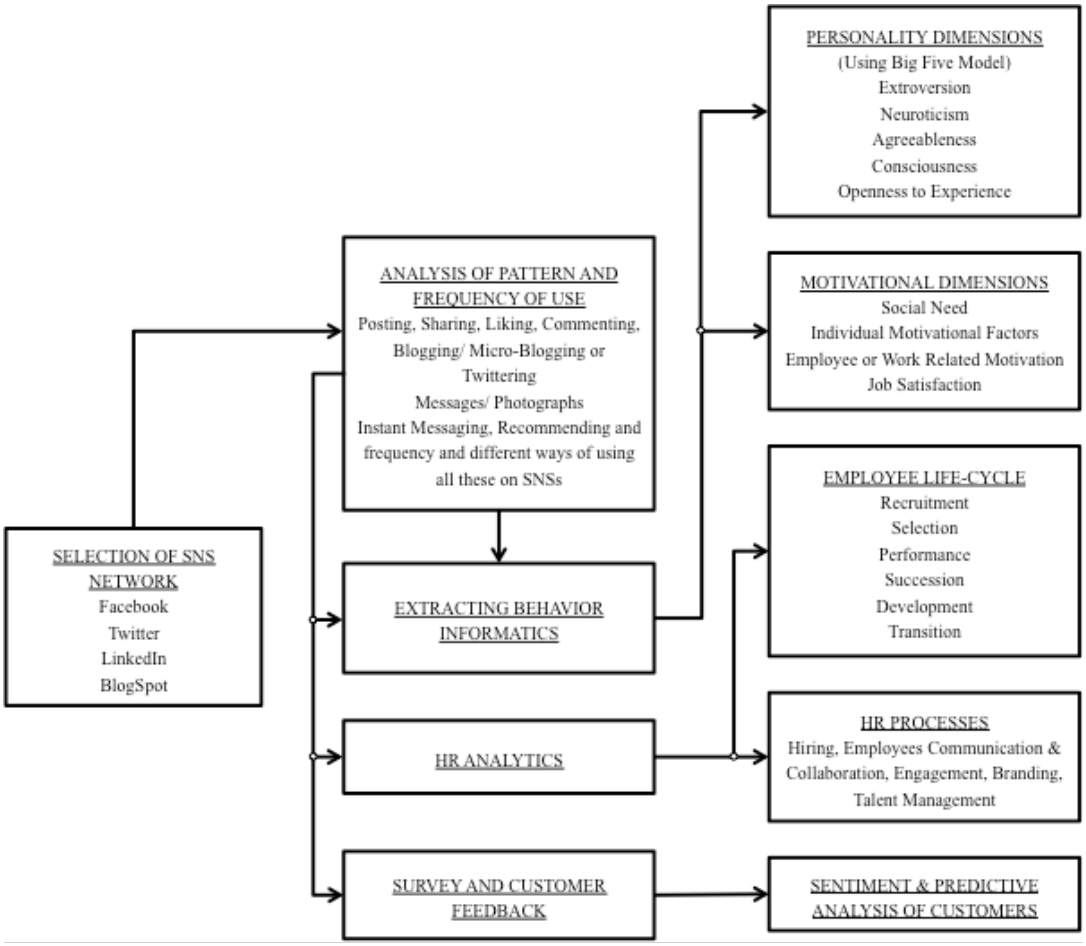


Based on these ideas, this framework includes two major components: sensing and seizing. Sensing is comprised of capturing and analyzing; capturing involves the use of social media monitoring tools to gather data from relevant social media sites. This data is then "analyzed" to generate valuable intelligence. This framework goes beyond SMA to include acting on this intelligence (which constitutes the seizing aspect of the framework).

2.4.2. The Sinha et al. Framework (2012)

Sinha et al. (2012) present a relatively comprehensive framework (Figure 2.3) that includes behavior analytics, human resources (HR) analytics, and customer analytics. This framework also focuses on providing SMA-related business benefits.

Figure 2.3 A Contemporary Model of SMA for Behavior Informatics, HR and Customers (Sinha et al., 2012)



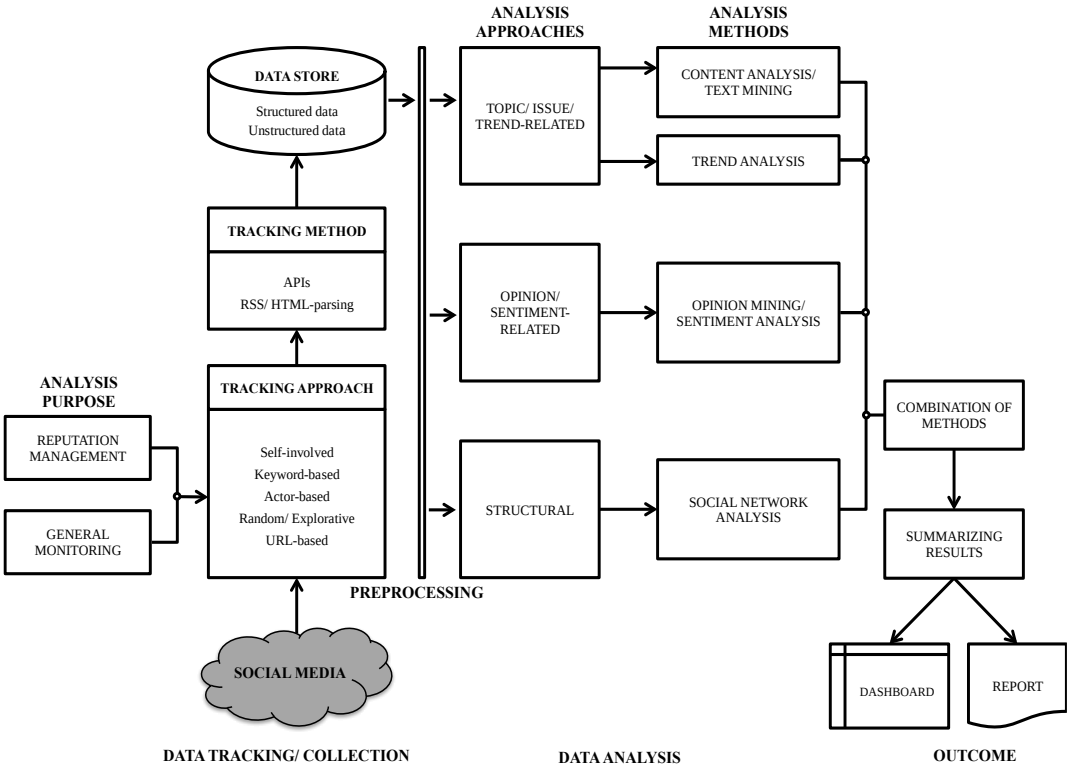
In this model, a firm starts extracting data from popular social media (e.g., Facebook, Twitter, LinkedIn, and BlogSpot). Later, the gathered data is analyzed based on relevant attributes (e.g., postings, likes, comments, and retweets). The goal, at this stage, is to extract, understand, and predict information associated with participant behavior, human resources, and customers. The behavior analytics module draws upon prior theoretical work in psychology (e.g., the five-factor model) to (i) relate the online behavior of participants on social media to their

personality traits and (ii) assess work-related motivational attributes of current employees (e.g., job satisfaction). The HR analytics module seeks to analyze the individual employee life cycle (from recruitment through retirement) and the management of other HR processes (e.g., hiring, retirement, employee engagement, and talent management). Furthermore, the customer analytics module is focused on sentiment analysis and predictive customer analytics to forecast future behaviors (e.g., purchases, churn, and spending). The module advocates the utilization of surveys to gather appropriate customer data from social media to facilitate customer analytics.

2.4.3. The Stieglitz & Dang-Xuan (2012) & The Stieglitz et al. Framework (2014)

The framework proposed by Stieglitz & Dang-Xuan (2012) differs from its previous model in that its primary focus is political analytics. In this framework (Figure 2.4), SMA-associated motivations are (i) reputation management and (ii) the general monitoring of the political climate by political actors and/or establishments. This framework provides a relatively detailed portrayal of data collection and analyses processes (vs. prior frameworks).

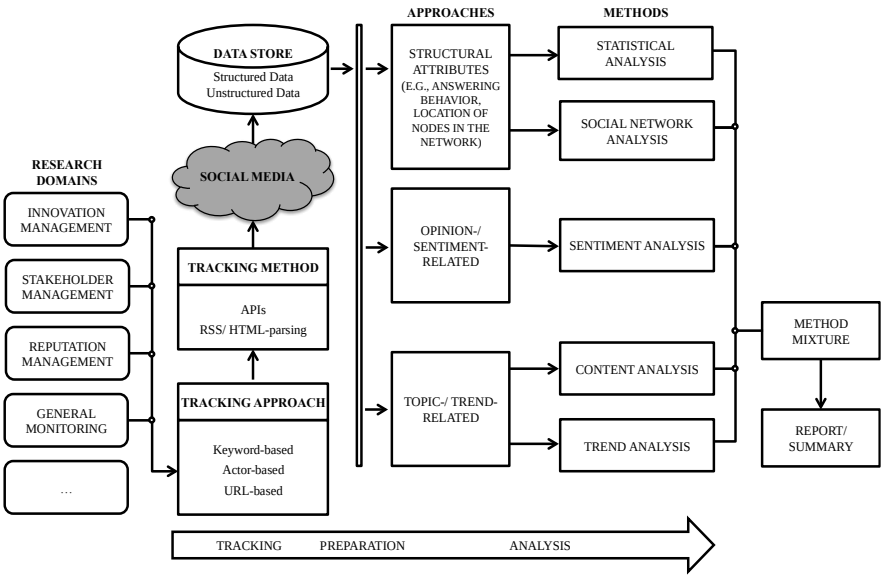
Figure 2.4 SMA Framework in Political Contexts (Stieglitz & Dang-Xuan, 2012)



This framework consists of a data tracking and collection module and a data analysis module, and each contains multiple sub-modules. The model presumes that relevant data resides in microblog (e.g., Twitter), social networking site (e.g., Facebook) and blog. Based on these different types of data sources, the authors provide different tracking methods, including search API and streaming API for Twitter, Graph API for Facebook, and. web crawling or tracking RSS feeds for blog. Between the data collection and tracking and analyses modules is the data pre-processing activity where data is cleansed and prepared for analysis.

The analysis module is concerned with different analysis approaches recommended for the gathered data. Analysis itself is aimed at the twin goals of reputation management and monitoring the political landscape. As with data tracking and collection, the authors articulate data analysis approaches (and methods) for each goal. The approaches for reputation management are the topic/issue/trend-related approach (using text mining and trend analysis as methods), the opinion/sentiment-related approach (using opinion mining and sentiment analysis), and the structural approach (using social network analysis). General monitoring is achieved using exploratory analysis of data collected using the exploratory/random approach. The specific analysis approaches and methods deployed for monitoring are similar to that for reputation management. Even though Stieglitz & Dang-Xuan (2012) focuses on political analytics, the framework may be extended for application in other business domains. Later on, Stieglitz et al. (2014) present an improved framework as depicted in Figure 2.5.

Figure 2.5 Social Media Analytics Framework (Stieglitz et al., 2014)

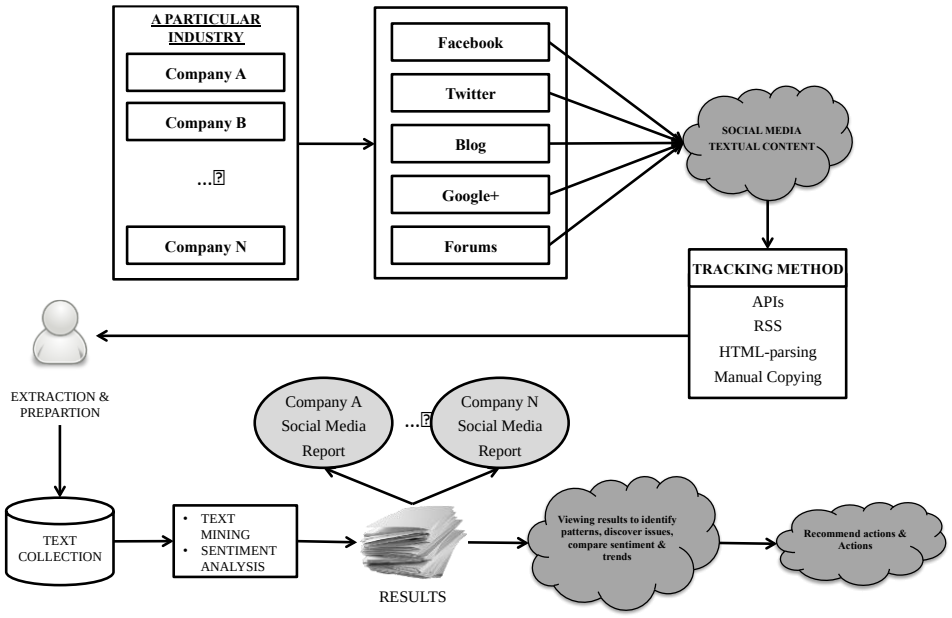


This model includes Innovation Management and Stakeholder Management as analysis goals in addition to Reputation Management and General Monitoring, and allows for other goals. The model also includes Statistical Analysis as part of the Structural approach to analyzing pre-processed data in addition to Social Network Analysis.

2.4.4. The He et al. Framework (2015)

Unlike previous models, He et al. (2015) propose a business SMA framework (Figure 2.6) focused entirely on Competitive Analytics. Social media data of competing firms in a particular industry (e.g., technology, banking) are extracted using available APIs crawling and parsing HTML, or even manual copying. Such data using various quantitative measurements (such as, number of fans/followers, number of posts, and frequency of posts) and qualitative metrics (such as, sentiments, emotions) from a competitor could be compared with a firm's own data. This result can provide the company the business intelligence to improve their competitive advantage. The authors contend that this data extraction activity is a constant process and the data is pre-processed to make it suitable for analytics efforts.

Figure 2.6 A Social Media Competitive Analytics framework with sentiment benchmarks for industry-specific marketing intelligence (He et al., 2015)



After the data cleansing, multiple analytics techniques including text mining, sentiment analysis, and social network analysis can be processed. The outcome can generate marketing

intelligence such as, “new knowledge (e.g., brand popularity) and interesting patterns, to benchmark industry sentiments and categories, to understand what their competitors are doing and how the industry is changing in various categories.” Such intelligence can then be used “to develop new products or services and to make informed strategic and operational decisions.”

2.5. A Framework of Social Media Analytics-based Decision Making

To date, the Stieglitz et al. (2014) framework is a relatively comprehensive frameworks reviewed here. It yet could be improved upon to provide a suitable framework that as especially tailored for business SMA. Based on the Simon’s Decision Making Model (Simon, 1960), this study designs a framework of SMA-based Decision Making (Figure 2.7) to provide benefits for business area. This study discusses the characteristics and concept of each step in the next couple paragraphs. The arrows in this framework (figure 2.7) represent the next possible step, and each

2.5.1. Analysis Goal(s)

Before starting to execute SMA, an organization should define its analysis-related goals. In Simon’s decision-making model, the “Intelligence” stage refers to searching and scanning the environment for conditions associated with problems and opportunities (Simon, 1960). Intelligence gathering identifies objective-related assumptions, motivations, and expectations (Nutt, 2007; Stapleton, 2003); involves scanning the environment; and provides information to (i) determine potential decision situations and (ii) formulate alternatives (Wally & Baum,1994). SMA examines the social media environment (to recognize problems and detect opportunities) and helps a firm improve its competitive advantage and performance via (i) designing a better business plan, (ii) improving business strategies, (iii) promoting successful product launches, and (iv) and developing new markets (Ahituv et al., 1998; Daft et al., 1988; Teo & Choo, 2001). Sense making means structuring the unknown into rationally accountable understandings to reduce confusion (Brown & Humphreys, 2003; Huber & Daft, 1987; Starbuck & Milliken, 1988; Weick, 1995) and is based vague questions, muddy answers, and negotiated agreements to develop a cognitive map of the current environment (Weick, 1993).

Thus, SMA involves the collection of information to help firms respond to their environments with meaningful and appropriate actions (Savolainen, 1993). For example, to develop a suitable marketing strategy, a firm needs to consider the opinions of customers (disseminated via social media networks). SMA also generates value insight, which goes beyond

sense making and provides actionable solutions. Insight Generation represents the knowledge of a firm (e.g., perspective, understanding, or deduction) and produces new information that yields actionable ideas (Cooper, 2006). Furthermore, the different goals of SMA can support a firm's decision-making processes in multiple ways. The main purpose of this SMA-based decision-making framework is to extract useful and valuable information to optimize a firm's decision-making abilities.

2.5.2. Social Media: Input Data

Prior to Intelligence Gathering, data must be collected via multiple social media sources (e.g., microblog, social network site (SNS), and online forum). This framework emphasizes that the data can be extracted from three domains: internal social media, external social media, and hybrid social media. Internal social media networks are social media portals utilized by firms. Firms now widely use social media (i) for communication purposes and (ii) to support other business-related needs. External social media refers to all the public social media networks on the Internet. Hybrid social media are social media portals that enable firms to interact with their customers through various (e.g., business-to-customer (B2C), customer-to-business (C2B), customer-to-customer (C2C)) methods. One example is the co-creation forum that provides the portal for a company to work with customers on generating ideas.

As with previous frameworks, multiple data-tracking approaches (e.g., API) can be applied in this stage. Because of the streaming nature of social media data, company requires to act effectively and efficiently when performing SMA. Advanced data query tools provide the capability to optimize the extraction of large amount of data flow continuously. For example, new data-tracking approach, namely Event Processor, detects and captures events, filters out noise, and monitor trends and correlations of data to support the SMA data query (Wootton, 2014). An Event Processor uses a dataflow architecture through the continuous query operators to preliminary filter out unnecessary information, improving the data extraction performance to provide instantly updated results (SAP, 2014). It helps to extract correlated, group, and aggregated summary social media data for stream analytics.

Stream analytics focuses on visualizing business in real-time, detecting urgent situation, and automating immediate actions (Gualtieri & Curran, 2014) to help company deal with

dynamic social media data stream. The “velocity” feature of Social Media data requires agile data analytics capability to provide real-time results (Best et al., 2012). The continuous querying and analyzing capability in stream analytics allows company to utilize social media to support real-time decision-making. On the other hand, data analyst can decide whether and which part of data should be stored for future reference. NoSQL database performs efficiently relative to traditional relational database in handling unstructured social media data (Leavitt, 2010). In SMA, NoSQL database provides the advantage of reading and writing data quickly, supporting massive data storage, easy to expand data storage size, and lower cost, compared to relational database (Han et al., 2011).

2.5.3. Intelligence

At this stage, this study incorporates concepts and tools, associated with information technology and business analytics, into the prior model to support the collection of valuable social media information. The velocity of social media data requires dynamic data analytics capabilities to provide real-time results (Best et al., 2012). Stream analytics focuses on visualizing business in real-time, detecting urgent situation, and automating immediate actions (Gualtieri & Curran, 2014) to help firms deal with dynamic streams of social media data. The continuous querying and analyzing capability enables a firm to utilize social media data to (i) improve agility of reaction and (ii) determine whether to adopt Stream Analytics and execute real-time data analysis. Stream analytics enables real-time analysis of social media data; however, this data can also be stored in a database and analyzed at a later time.

The next step of intelligence gathering involves preprocessing social media data to ensure the data are readable and reliable; aforementioned issues (e.g., unstructured social media data) make it difficult to extract accurate information (Mayeh et al., 2012). Furthermore, social media data often contain unclear sentiments, which hinder further usage of the data (Mosley, 2012). Thus, prior to applying SMA to social media data, preprocessing is required to ensure data validity. This study offers a couple of suggestions for dealing with this issue. For example, the stop-word approach can be employed to preprocess social media data within text formats. This approach uses a preset wordlist to filter unnecessary words (e.g., “the,” ”is,” and ”a”) and only keeps words with meaningful sentiments. However, in order to preprocess network data, the boundary of the network, for instance, must be defined (e.g., by time period, community size, demography, or other criteria based on needs). Preprocessing activity data quantifies specific user behaviors (e.g., the frequency of posting content, replying to comments, or providing feedback).

After preprocessing social media data, multiple analytics approaches can be executed. This dissertation identifies the most common analytics techniques in this framework.

- Trend Analysis (Predictive Modeling): Trend analysis provides up-to-date information on trend events and thus supports firms that are seeking to (i) develop or improve their business strategies (Schaust et al., 2013), (ii) understand market trends, and (iii) react dynamically. It is used to "predict future outcomes and behaviors based on historical data collected over time." (Fan & Gordon, 2014, p. 78). This dissertation argues that Predictive Modeling should be part of Trend Analysis, used to estimate uncertain events in the past to make future prediction. From a marketing perspective, SMA can measure the outcomes of campaigns via knowledge of WOM effects. For example, Colbaugh & Glass (2011) have developed a mechanism to catch potential social media trends and identify how people react to products and communicate about them via social media. Real-time, streaming social media data can provide first-hand information and support crisis management.
- Topic Modeling: SMA also extracts the main topic people are talking about (or predicts it). Topic modeling refers to the technique that looks for patterns in the use of words to discover the hidden semantic structure in large archives of documents or text corpus. It also injects semantic meaning into vocabulary, in which a "topic" consists of a cluster of words that frequently occur together (Wang et al., 2013a; Wang et al., 2013b). This approach can capture specific, dominant themes (and topics) from a vast amount of text content. SMA, via the utilization of diverse statistics and machine-learning tools, can extract the most popular topics, improve business strategies (Fan & Gordon, 2014), and enable firms to react immediately (e.g., on product improvements or service complaints). Wang et al., 2013 have designed a new algorithm to identify experts in online communities who support firms locating such information. Also, Wu & Lin (2012) have designed a mechanism to help users explore topics and organize them in Wikipedia accurately. This provides users with decision support as they search online resources.
- Sentiment Analysis/Opinion Mining: Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to extract subjectivity and polarity from text (potentially also speech), and semantic orientation refers to the polarity and strength of words, phrases, or texts (Li et al., 2014; Taboada et al., 2011). Firms can also monitor the sentiments of social media data to understand the attitudes of their customers (about their products and services). SMA tools sort through user-

generated content (e.g., postings, blogs, and tweets) to help determine whether opinions contain positive or negative attitudes. For example, Vorvoreanu et al. (2013) have studied the public opinion about Indianapolis, in the context of Super Bowl XLVI. They emphasize that sentiment analysis enables firms to determine outcomes from social media postings. Since user-generated content can be considered as objective description, a firm can accurately identify public sentiment (vs. soliciting opinions), which may cause bias in some circumstances. Also, Yang et al. (2011) emphasize that machine learning and semantic-oriented approaches should be combined to train the model. This can improve sentiment identification and generate a more accurate result.

- Social Network Analysis: Social media involves users sharing information and content via online social networks. Thus, one of the main issues is to understand the structure of social networks for estimating information-diffusion speeds and WOM effects. Social network analysis refers to the approach based on graph-theoretic properties to characterize structures, positions, and dyadic properties (such as the cohesion or connectedness of the structure) and the overall "shape" (i.e., distribution) of ties (Borgatti et al., 2009). SMA, for instance, uses social network analysis to identify the "opinion leader" and thus improve analytics performance. Furthermore, Boden et al. (2013) argue that the significant quantities of data on social media networks can affect SMA performance. They support the identification of key nodes, at the initiation of data collection, to improve SMA performance. In the model of Boden et al. (2013), they further emphasize that opinion leaders can be identified via (i) analysis of users' postings within specific timeframes and (ii) ties between users.

In this framework, the multiple approaches discussed above could be applied parallel or sequentially. For example, one may rune Social Network Analysis first to identify a small group of people and then run Topic Modeling to identify important subjects. On the other hand, one could integrate the results of Social Network Analysis and those of Topic Modeling to synthesize a different outcome. Thus, SMA monitors and evaluates opinions from social media and can activate the decision-making process (Engel et al., 1978; Fletcher, 1988). Furthermore, ideas detected via social media can support (i) unfulfilled market needs or (ii) solutions that satisfy a need (O'Connor & Rice, 2001). Problem recognition is the process to perceive the difference between the "ideal state of affairs" and the "actual situation." (Engel et al., 1978; Fletcher, 1988) It is critical for the "effective management of complex, real-world situation." (Klein et al., 2005) Opportunity detection refers to the match between an unfulfilled market need and a solution

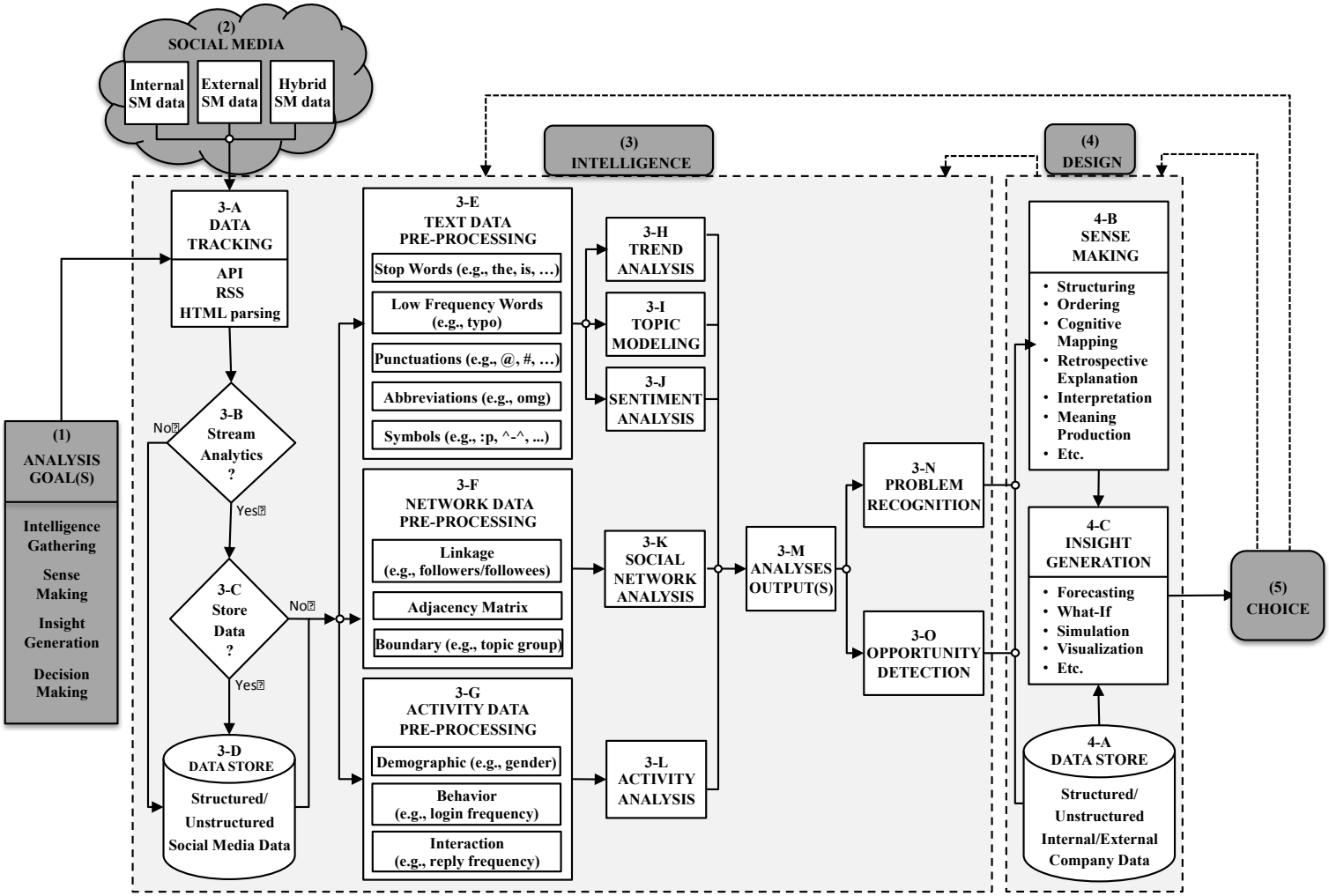
that satisfies the need (O'connor & Rice, 2001). It identifies connections between "breakthrough ideas" and "initial innovation evaluation processes." The ideas detected from social media may support an unfulfilled market need or a solution that satisfies the need (O'Connor & Rice, 2001). As noted earlier, SMA takes the approach of "listening" to users, rather than "asking" for user input.

2.5.4. Design and Choice

In the design stage, business SMA supports sense making, and/or insight generation as firms "devise courses of action aimed at changing existing situation(s) into preferred ones" (Simon, 1996, p. 130). Sense making here is defined as structuring the unknown into sensible, "sensable" events in their efforts (Brown & Humphreys, 2003) and insight generation refers to going beyond knowing what the solution is for a given set of input data and discern why the solution is what it is (Steiger, 1998). SMA outputs structure the unknown truth (or unclear understanding) to make sense of current environments. Structured truths (e.g., time, space, movement, step taking, situation, and outcome) create order and develop a cognitive map of the environment to help a firm respond with meaningful and appropriate actions (Dervin, 1998; Ring & Rands, 1989; Savolainen, 1993). After a firm makes sense of an environment, it can contemplate multiple "what-if" scenarios and generate new insights (Geoffrion, 1976; Steiger, 1998).

SMA outputs, combined with structured and unstructured firm data, can also generate insights to derive meaningful ideas, directions, solutions, and recommendations associated with decision designs (Heinrichs & Lim, 2005). These structured and unstructured firm data could be came from external environment or internal company data. For example, one can integrate the SMA outputs such as customer satisfaction with actual sales data to generate insights. Finally, the utilization of SMA capability can provide relatively comprehensive support for decision makers. In the Choice stage, decision makers assess anticipated problems and solutions before execution, the entire decision-making process is continuous, and whenever something needs to be adjusted or addressed, firms make related changes to improve outcomes.

Figure 2.7 A Framework of SMA-based Decision Making



2.6. Concluding Remark

In this chapter, this study reviews the literature related to SMA. Drawing on (and augmenting) available SMA definitions, this dissertation develops an integrated, unifying definition of business SMA, with a view toward providing a nuanced starting point for future business SMA research. This definition goes beyond (i) a customer focus (to encompass external and internal organizational environs) and (ii) intelligence gathering to accommodate such business activities as sense making; insight generation; problem and solution detection and exploitation; and decision-making.

This dissertation also identifies several benefits of business SMA. This dissertation identifies and categorizes several challenges facing business SMA today, along with supporting evidence from the literature (that documents such challenges and offers mitigating solutions in particular contexts. Still, comprehensive, viable solutions to several of these challenges remain elusive.

This dissertation furthers a conceptual understanding of business SMA (and its many aspects), grounded in recent empirical work and is a basis for future research. Finally, this dissertation reviews previous SMA frameworks from the literature to provide a clear roadmap for framework development. Based on the Simon's decision-making model, previous studies, and current SMA approaches, this dissertation presents a relatively comprehensive framework for SMA-based decision making. This framework gives researchers a structure for design and interpretation of their own SMA investigations, and help practitioners to develop their own SMA system/tool/software.

Chapter 3. An Assessment of Algorithmic Social Influencer Identification Approaches

3.1. Introduction

Social media has become one of the most popular online services, and as a consequence, it includes vast amounts of information flow and data that have yet to be completely explored and analyzed. The previous chapter emphasized the potential of SMA. Given the nascent state of SMA research, there remain many unanswered questions that have attracted my research interest. This chapter focuses on one type of SMA implementation, namely, social influencer identification. This study reviews existing approaches to influencer identification, which reflect the framework of SMA-based decision-making processes and illustrate the importance of developing SMA, and multiple experiments are conducted to show how SMA applies to a real-world problem and provides decision support.

Influencer identification in SMA supports business decision-making in many ways. For example, Subramani & Rajagopalan (2003) identify that social influencer plays a crucial role for viral marking in (i) its recommender role to passive or actively persuade people to adopt new product and (ii) the level of network externalities for expanding current customer base. Goodman et al., (2011) also emphasize the importance of social influencers in increasing company brand awareness through social media. Moreover, company can promote its product by recruiting these influencers to share user experience and product-related information. These influencers also provide valuable feedbacks based on their important roles in the market. Identifying social influencers from social media provides company a niche to explore its territory and, at the same time, explore its market. Hence, this dissertation focuses on social influencer identification by solving current issues of influencer identification approaches and provides a relative comprehensive understanding of these approaches for helping companies to support their business activities.

In implementing SMA for identifying influencers, the first issue that must be addressed is the complexity of the social media data. As social media is continuously accumulating vast amounts of structured and unstructured data, the four Vs that affect data analytics—volume, velocity, variety, and veracity—are different in social media than in other data analytics areas. First, the volume of social media data is rapidly increasing and thus requires a relatively efficient and effective method to analyze it. Second, because of its streaming nature, social media has a high velocity of data production. The critical issue is to provide a flexible and expeditious system

to handle this structured and unstructured data. The combination of different types of data (i.e., text content, user activity, and social network) in social media data mean a greater variety of data and thus more complexity that must be addressed in order to extract valuable information from the data. At the same time, this makes it more challenging for analysts to prove the veracity of the data analytics results. Hence, to solve these problems, this dissertation examines the assessment of different sizes of data inputs by different analytics approaches to identifying social media influencers. Based on the experiment designs, this study develops a relatively comprehensive theoretical foundation for understanding SMA implementations for influencer identification.

In this chapter, this study first describes existing explanations and categorizations of social media in order to provide a clear understanding of current developments in social media research. Next, this study discusses the literature related to social influencer identification and the use of SMA in identifying influencers in social media network. Scholars in many research areas have developed different approaches to conducting SMA and identifying social influencers. The mechanisms, which include methods from statistics, computer science, and mathematics, are applied to different categories of social media and employ different types of data. However, there are no clear rules about which of the various algorithms/methods should be employed when analyzing different types of social media or which kinds of data should serve as input. Hence, this study categorizes the approaches found in the literature and design multidimensional experiments to illustrate the advantages and weaknesses of the different approaches to identifying social influencers. By examining different kinds of input data, this research provides a relatively comprehensive view of SMA implementations for influencer identification.

This chapter starts with a review of social media literature related to social media and categorization. The following sections discuss the concepts related to influencer identification as well as how SMA is applied using different algorithms/methods in this area. Then the section describes the experiment design for this research and discusses the results of the experiments. The last section presents conclusions about this comparative assessment of approaches to influencer identification.

3.2. Social Media Development

Since the emergence of social media, different types of services have been introduced to the market and accumulated users at a dramatic speed. To date, social media has attracted more than two billion people, or more than 30% of active Internet users globally (Regan, 2015).

Generally speaking, social media provides services on Web 2.0 portals to support interactions between individuals and communities, allowing them to produce and share content. The use of social media has extended beyond the individual level and is becoming more and more appreciated by companies. The flourishing of social media has substantially changed how individuals, communities, and organizations communicate and interact within one another (Ngai et al., 2015). Therefore, scholars and practitioners have devoted great effort to providing a more comprehensive understanding of social media for business purposes.

Zeng et al. (2010, p. 13) briefly define social media as "a conversational, distributed mode of content generation, dissemination, and communication among communities." Kaplan & Haenlein (2010, p. 61) further clarify social media as "a group of Internet-based applications that build on the ideological and technological foundations of web 2.0, and that allow the creation and exchange of user generated content," and Weinberg & Pehlivan (2011) emphasize that social media is the group of applications tools that provide innovative services on Web 2.0 computer-based platforms. Kietzmann et al. (2011, p. 241) provide a relatively precise explanation: "social media employs mobile and web-based technologies to create highly interactive platforms via which individuals and communities share, co-create, discuss, and modify user-generated content." This study adopts this definition and categorizes social media types as explained in the next section.

3.3. Social Media Categorization

Based on Kietzmann et al.'s (2011) definition of social media, social media is the services provided on Web 2.0 portals to support interactions between individuals and communities, allowing users to produce and share content. Since the emergence of social media, different types of service have been introduced to the market Mangold & Faulds (2009) categorize social media as shown in Table 3.1:

Table 3.1 Categorization of Social Media (Mangold & Faulds, 2009)

Type	**Example**
Social Networking Site (SNSs)	MySpace, Facebook
User-sponsored Blog	The Unofficial Apple Weblog, Cnet.com
Company-sponsored Website/Blog	Apple.com, P&G's Vocalpoint
Company-sponsored Cause/Help Site	Dove's Campaign for Real Beauty, click2quit.com
Invitation-only Social Network	ASmallWorld.net
Business Networking Site	LinkedIn
Collaborative Website	Wikipedia
Virtual World	Second Life
Commerce Community	EBay, Amazon.com, Craig's List, iStockphoto, Threadless.com
Podcasts	"For Immediate Release: The Hobson and Holtz Report"
News Delivery Site	Current TV
Educational Materials Sharing	MIT OpenCourseWare, MERLOT
Open Source Software Community	Mozilla's spreadfirefox.com, Linux.org
Social Bookmarking Site	Digg, del.icio.us, Mixx it, Reddit
Creative Works Sharing Site	
1. Video Sharing Site	YouTube
2. Photo Sharing Site	Flickr
3. Music Sharing Site	Jamendo.com
4. Content Sharing combined with assistance	Piczo.com
5. General Intellectual Property Sharing Site	Creative Commons

This categorization is relatively confusing and at odds with the common formats of social media. Over the past few years, some of these services have become ever more popular but others are just like the dot-com bubble, that is, they no longer exist. This categorization needs to be revised to provide a more up-to-date understanding of the categories of social media. Sterne (2010) defined the categories of social media to include:

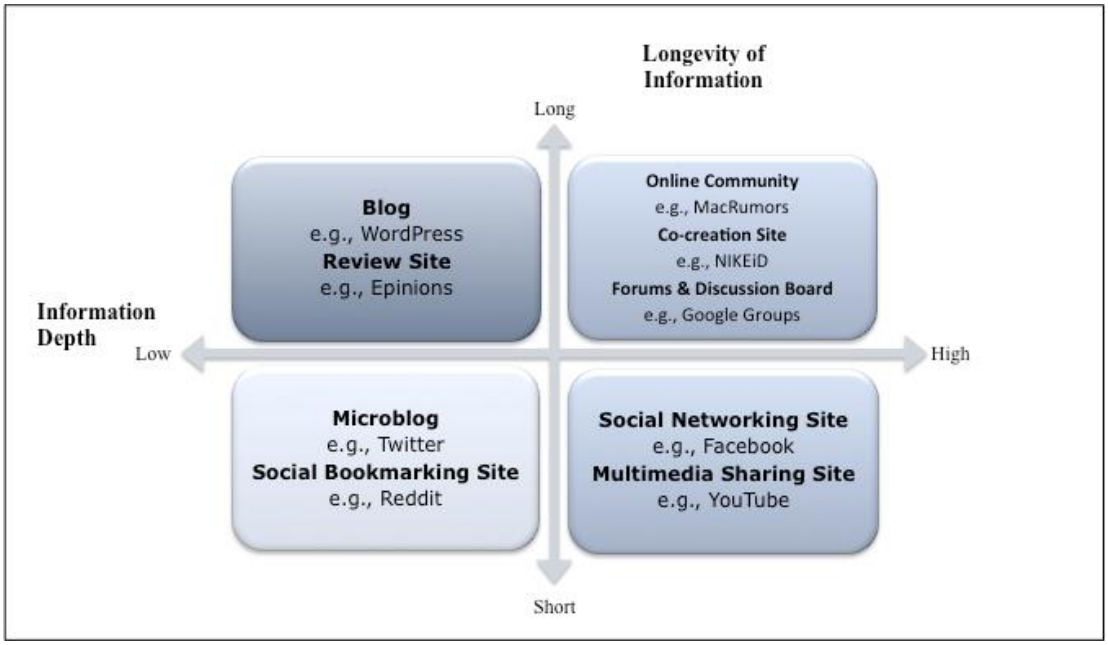
1. forum and message board,
2. review and opinion site,
3. social network,
4. blogging,
5. microblogging,
6. bookmarking, and
7. media sharing.

Hoffman & Fodor (2010) identified social media categories as follows:

1. blog,
2. microblog (e.g., Twitter),
3. co-creation site (e.g., NIKEiD),
4. social bookmarking site (StumbleUpon),
5. forum and discussion board (e.g., Google Groups),
6. review site (e.g., Yelp),
7. social networking site (e.g., Facebook, LinkedIn), and
8. multimedia sharing site (e.g., Flickr, YouTube).

This dissertation adopts Hoffman & Fodor's (2010) categorization, which this study covers the most common social media types. Because of the different features of these different types of social media, Weinberg & Pehlivan (2011) provided a figure to illustrate the different information depths and half-lives of different social media types. This figure clearly depicts the characteristics of different types of social media. Combining the social media types from Hoffman & Fodor (2010) with Weinberg & Pehlivan's (2011) figure results in an updated figure providing a different view of social media categorization (Figure 3.1).

Figure 3.1 The Half-life of Information and Information Depth of Different Social Media Types (Adapted from Weinberg & Pehlivan, 2011 & Hoffman & Fodor, 2010)



Based on this categorization, this dissertation synthesizes the definitions of each type of social media from the literature, documents specific features of each type, and emphasizes their data structures. For example, the contents of blogs are normally long articles with a strong text data structure. The network between users in a blogosphere is relatively weak, and most of the social relationships are bilateral (writer and readers) but not multidirectional. User activity is relatively moderate compared to micro-blogs (e.g., Twitter), but not as weak as on co-creation platforms (e.g., BMW Group Co-Creation Lab), which mainly depend an active co-creation project. Table 3.2 summarizes these characteristics of each type of social media:

Table 3.2 Definition and Features of Different Types of Social Media

Social Media Type	Definition	Features	Data Structure	References
Blog (e.g., Blogger)	A web-based publishing tool which consists of a series of posts by the author(s) on a personalized web page, with posts usually arranged in reverse chronology from the most recent post at the top of the page.	• New entries at the top, updated frequently • Interaction in subscribing, commenting, citing contents • Bloggers with readers • A complex social network often called a blogosphere	• Text: Strong • Network: Weak • Activity: Moderate	(Blood, 2002; Chau & Xu, 2012; Lin & Kao, 2010; Minocha & Roberts, 2008)
Microblog (e.g., Twitter)	A new form of communication in which users can describe their current status in short posts distributed by instant messages, mobile phones, email or the Web.	• Broadcast and share update of user's activities, opinions, and status • Real-time updates • Flexibility of access (e.g., mobile devices) • Lightweight architecture (e.g., word limited to short message)	• Text: Moderate • Network: Strong • Activity: Strong	(Honey & Herring, 2009; Java et al., 2007)

Social Bookmarking Site (e.g., StumbleUpon)	A system allows users to share their tags for particular resources. In addition, each tag serves as a link to additional resources tagged the same way by others.	• Store, manage, search, organize, and share bookmarks online • Self-assigned or selected tag to bookmarks • Search bookmarks by individual or keyword	• Text: Weak • Network: Strong • Activity: Strong	(Barnes, 2011; Gray et al., 2011; Marlow et al., 2006)
Social Networking Site (e.g., Facebook, LinkedIn)	A web-based service that allows individuals to (1) construct a public or semipublic profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system.	• Visible personal digital profile • Display an articulated list of friends • Public display of connections • Bridge online and offline relational connection • Provide public search and private access	• Text: Weak/ Moderate • Network: Strong • Activity: Strong	(Ellison, 2007; Kane et al., 2014)
Review Site (e.g., Yelp)	A site that provides peer-generated product evaluations posted on company or third-party websites.	• Interaction in writing reviews, rating products or brands, and forwarding comments • Numerical star ratings • Open-ended customer-authored comments	• Text: Moderate • Network: Weak • Activity: Weak/ Moderate	(Chevalier & Mayzlin, 2006; Hennig-Thurau et al., 2003; Libai et al., 2010; Mudambi & Schuff, 2010; Munzel & H. Kunz, 2014)

Multimedia Sharing Site (e.g., Flickr, YouTube).	A channel allows users to display content that they uploaded; videos from other members; videos favorited by the channel, their friends, and subscribers; as well as channels that they subscribe to.	• Personalized page or channel • Upload and share multimedia contents • Creators are also consumers	• Text: Weak • Network: Moderate/ Strong • Activity: Moderate/ Strong	(Raymond, 1999; Susarla et al., 2012; X. Zeng & Wei, 2013)
Co-creation Platform (e.g., NIKEiD)	A platform that provides the paticipation for users along with producers in the creation of value in the marketplace.	• Customers create and construct value • User experiences with resources, processes and contexts • Company and customers have specified roles and goals	• Text: Moderate • Network: Weak • Activity: Weak	(Grönroos & Voima, 2013; Zwass, 2010)
Community, Forum and Discussion Board (e.g., Google groups)	A portal that users can use to discuss mutiple subjuects and topics based on personal interests and form communities and groups.	• Users discuss specific topics with reliable information, consumer relevant contents, and strong influence • Text-based discussion	• Text: Strong • Network: Strong • Activity: Strong	(Bickart & Schindler, 2001; Marett & Joshi, 2009)

3.4. Social Influencers

This study concentrates mainly on identifying the important individual(s) in a social media network for business decision support. These important users have relatively strong social influence in different types of social media. Since the 1940s, people have been interested in the

influential individual(s) in specific social groups. These individuals, who have been called "opinion leaders," are "certain people who are most concerned about the issues as well as most articulate about it" (Lazarsfeld et al., 1944, p. 49). The early literature focused on people who have strong opinions and are politically influential with respect to their relatives. The definition of "opinion leader" was later extended to have a broader scope. Katz & Lazarsfeld (1955) explain that these "opinion leaders" are people who are more influential than others within their social networks. They consider themselves experts in a specific area of interest (e.g., home policies or fashion) and are asked for advice in that area. Rogers & Cartano (1962) also emphasize that opinion leaders are "individuals who exert an unequal amount of influence on the decisions of others" (p. 435).

The idea of an opinion leader involves more than simple informal advice seeking from peers. The opinion leader dominates attitudes or behavior in his/her social network and has a strong influence on the decisions of others (Black, 1982; King & Summers, 1970). Chan & Misra (1990) further point out that opinion leaders produce greater knowledge about and interest in a particular product or issue than do others. Recently, opinion leaders have also been called "influencers" (Torres et al., 2016), "social influencers" (Langner et al., 2013), and "leading users" (Yi-si & Guo-xin, 2012). This dissertation uses the term "influencer" to define these individuals because this term not only points to the power of these individuals but also emphasizes that their influence is based on a social media network. The social media influencer who conveys ideas to others through social media has considerable impact.

3.5. Influencer Identification

Traditionally, marketing scholars have mainly used a survey methodology to identify the influencers in a small social group/network. For example, Lazarsfeld et al. (1944) use a self-report measure questionnaire to ask participants about their media usage, activity, and interest in election campaigns. Based on user behavior, this study argues that the influencer uses the media more frequently and has a stronger interest than others. Katz & Lazarsfeld (1955) subsequently combine the self-report measure and third-party ratings. After filling out self-report questionnaires, participants were interviewed and asked to identify people from whom they (the participants) sought advice. Based on this mixed measurement, the authors identify the influencers as those individuals who actively used the media and gave most of the advice. Summers (1970) reviews previous research and develops a questionnaire combining three different dimensions—demographic characteristics, social and attitudinal characteristics, and

topic-oriented characteristics—to measure 1,000 homemakers in order to identify the individual with the highest score as the influencer. In contrast, Schenk & Rössler (1997) adopt a personality strength scale and use social network analysis to identify influencers among 900 adults in Germany. However, the manual approach is not scalable when considering a large group of users in social media (Hudli et al., 2012). Researchers have since adopted different SMA approaches to analyze social media users in order to identify the influencers in huge social media networks.

SMA research categorizes social media data into three different types: text content, user activity, and social network. Text content is the main production of social media activity. The different kinds of text content range from content posts to comments/replies to feedback, tags, and even titles. These text contents influence readers and create an impact with the conveyed information. Social media user activity includes information about different user behaviors: click-through data, login frequencies, and the number of comments, feedbacks, and replies produced by a user. For example, an individual who uses a specific social media platform more often than others may attract more followers because he/she produces content or interacts with others more frequently. Social media allows individuals to connect in multiple virtual ways, and such connections weave a network of users who are socially interacting with each other. The individual who actively replies to comments, provides feedback, and networks with other users accumulates more social network connections. These three different types of data provide SMA with a good resource for analyzing users and finding the most influential one(s).

For analyzing the different types of social media data, researchers have adopted alternative approaches to influencer identification. For example, the development of text mining and sentiment analysis techniques have facilitated current SMA researchers' application of these methods to analyze individual opinions in social media posts (Khan et al., 2014; Li et al., 2014). Previous literature has emphasized that the text content an individual produces will have an impact on his/her social influence and position in social media. Social network analysis has also been employed in this area. This group of researchers claim that an individual with high centrality (i.e., more incoming/outgoing network ties) will spread more information to the whole network than others, and thus become the influencer (Borgatti, 2005). Activity analysis is also used to evaluate how an individual impacts a network (Butler, 2001). An individual's activity affects his/her social structures in that it "facilitates information exchange, influences social behavior, and even draws new users into the fold" (Huffaker, 2010, p. 595).

Some research has explored using SMA to identify social media influencers. For example, Huffaker (2010) applies hierarchical linear modeling (HLM) to identify influencers based on message reply triggering, conversation sparking, and language diffusing. He argues that communication activity, the social network, and language use can be used to measure the scale of individuals' influence. Li & Du (2011) designed a model to identify social influencers in online social blogs. The authors use four elements in their model—Blog content, Author properties, Reader properties, and Relationship (BARR)—as measurements to identify influencers. They build an ontology model for a marketing product to identify hot topics in the social blog, and then they locate the influencer based on discussion of these hot topics. Susarla et al. (2012) studied YouTube to identify the most influential channel in the multimedia-sharing social media. The authors investigated the social network structure and properties to examine the information diffusion model. Li et al. (2013) employed negative feedback and used a two-step approach to identify influencers. The first step creates a list of candidate influencers using a supporting vector machine (SVM) approach, and the second step forms the final list of social influencers by filtering out negative feedback.

Through the explosive growth in SMA, previous research has generated increasing interest in finding automated ways to discover social influencers (i.e., opinion leaders) in various social media settings. The literature shows that beginning in 2008, more than 20 variants of 6 basic approaches have been proposed. Yet there is no comprehensive study investigating the relative efficacy of these methods in specific settings. The next subsection reviews the literature to categorize current approaches to identifying social media influencers.

3.6. Use of SMA in Identifying Influencers

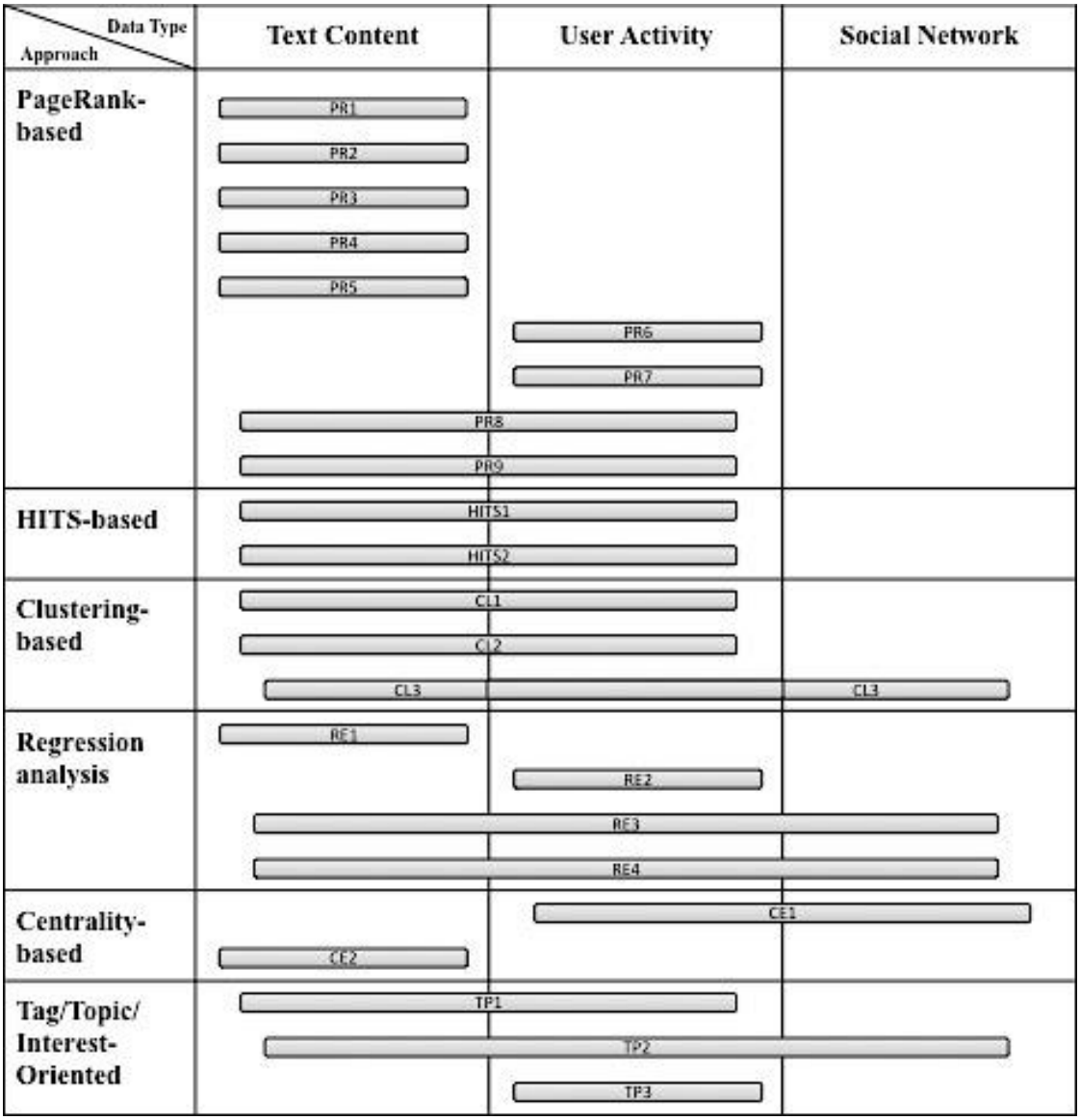
This study began by reviewing the literature, focusing on identification of social influencers that uses non-survey approaches. Google Scholar is used to search multiple journals (*MIS Quarterly*, *Information Systems Research*, *Journal of Management Information Systems*, *Journal of the Association for Information Systems*, *Decision Support Systems*, *Information and Management*, and *Management Sciences*) for the keyword "opinion leader(s)" appearing anywhere in an article and the keywords "opinion leader(s)," "influencer(s)," "leader(s)," and "leadership" appearing only in the title. This study also used the same search terms to find papers in the ACM Digital Library and IEEE Xplore Digital Library.

After reviewing resulting papers, this study finds three main data sources used in SMA to identify social influencers in social media. In addition, more than 20 variants of 6 basic approaches have been proposed. The basic approaches are:

1. PageRank-based algorithms,
2. hyperlinks-induced topic search (HITS)-based algorithms,
3. clustering-based algorithms,
4. regression analysis,
5. centrality measurement, and
6. tag/topic/interest-oriented algorithms.

Figure 3.2 codes and categorizes the identified research to illustrate the SMA approaches that are used with different types of input data. For example, the code “PR8” represents a PageRank-based algorithm using text and activity data as input data. The PageRank algorithm is a linkage-based algorithm that uses the connections between nodes to measure the importance of each node in an overall network. The authors who use the PR8 approach adopt PageRank as the base algorithm and modify it to blend text content and user activity with the linkage between nodes to evaluate the influence rank of each node. Nodes with higher ranks represent a higher level of social influence than nodes with lower ranks. Next, this research briefly explains the fundamental ideas behind each of the different approaches to SMA. A more completed discussion of each algorithm/ method appears in Appendix I.

Figure 3.2 Social Influencers Identification Approaches and Input Data Types



3.6.1. PageRank Algorithm

The PageRank algorithm was proposed by Page et al. (1999) to analyze the structure of the linkage between pages on the World Wide Web (WWW). In the Internet, each page has hyperlinks that link forward to other pages. Based on the incoming and outgoing links, these pages connect to each other as a network. Page et al. (1999) argued that the linkage between pages could be the main indicator for measuring the importance of each webpage. The PageRank algorithm supports a web search engine's effort to locate the most important pages given an input keyword. It iteratively calculates links to measure the importance of each page. A page with a

larger number of incoming links is more important in its network and has a higher rank. The simple version of the PageRank algorithm is as follows:

$$\boldsymbol{PR(i) = c \sum_{j \in B_i} \frac{PR(j)}{N_j}}$$

Where B_i is the set of pages that point to page *i*, N_j is the total number of outgoing links in page *j*, $PR\ (j)$ is the PageRank value (importance) of page *j*, and *c* is a factor used for normalization to make sure that the PageRank values of all pages are comparable (Haveliwala, 2002). The basic idea is that each page confers specific units of rank on others, and the summation of these ranks represents the importance of a page. The algorithm will continue the iteration until the rank is stable.

3.6.2. HITS Algorithm

The well-known HITS algorithm was designed by Kleinberg (1999). The original purpose of this algorithm was to analyze linkage structure in the WWW to support information extraction relevant to a specific topic from a collection of web pages. Through an iterative process, the HITS algorithm identifies the most "authoritative" pages depending on the linkage structure in a specific topic space (Kleinberg, 1999). To analyze the linkage structure, a collection of webpages is defined as a directed network G = (V, E), where V is the set of nodes representing webpages, and E is the set of edges representing the linkages between pages. If there is a hyperlink in page *i* that points to page *j*, then there is a directed edge from *i* to *j* in the network G. The HITS algorithm has two steps that are executed sequentially. The first step is a sampling stage, which narrows the original network down to a reasonably sized subnetwork, whose nodes are highly relevant to the search query. In the original development of the HITS algorithm, Kleinberg (1999) used search engine results to identify 200 sample nodes that were highly relevant to the search query topic.

The second step is a weight-propagation step. In the original HITS algorithm, this step calculated the degree of a node *v* by measuring the total number of nodes that point to it and that it points to. Each node is given a non-negative *authority weight a* (incoming links), and a nonnegative *hub weight h* (outgoing links). During the iterative process, the authority weight and hub weight of each node will be maintained and updated. If a node *i* is pointed to by many nodes

with a high hub weight, its authority weight will be high. In the other words, for node *i*, the value of a_i is the sum of h_j over all pages *j* that link to *i*:

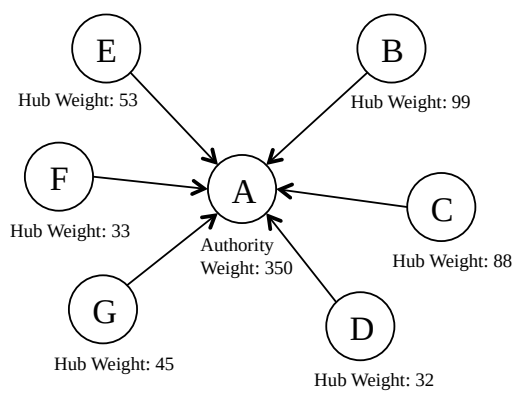
$$\boldsymbol{a_i} = \sum_{\boldsymbol{j:j \to i}} \boldsymbol{h_j}$$

At same time, if node *i* also points to many nodes with high authority weight, its hub weight h_i will be also high. The value of h_i is the sum of a_j over all pages *j* that *i* links to:

$$\boldsymbol{h_i} = \sum_{\boldsymbol{j:\ i \to j}} \boldsymbol{a_j}$$

Ultimately, both the authority weight and hub weight will achieve convergence and the actual authority nodes with the highest values of authority weight will be identified (Kleinberg et al., 1999). In Figure 3.3, the most authoritative node A with the highest authority weight has many incoming links from the hub nodes with high hub weight.

Figure 3.3 Illustration of HITS Algorithm



3.6.3. Clustering Algorithm

Another group of researchers adopted a clustering algorithm for social influencer identification. A clustering algorithm is an unsupervised data classification method; these algorithms have been applied to many research contexts in many different areas. The basic idea of clustering algorithms is to use predefined features of the data to classify observations into different clusters (Jain et al., 1999). To use clustering algorithms for social influencer identification, researchers have determined the characteristics of social media influencers and

adopted these as the data features for clustering users. Based on these predefined data attributes, users with similar rankings will be clustered into the same group, and ultimately, all of the social influencers will be in the same cluster.

3.6.4. Regression Analysis

Regression has also been adopted for influencer identification. The main purpose of the researchers who use regression analysis is to measure the correlation between the factors used to identify influencers and the metrics used to evaluate the influence these people may cause. For example, Huffaker (2010) regresses online social influence with three different metrics. He measures the communication activity of online users, their social networks, and the language usage in their online posts to evaluate the social influence of each user in Google Forum. He then regresses the social influence of the users with the capabilities of triggering replies, sparking conversation, and diffusing language. This group of research papers focuses on how social influencers can cause different manners of impacts.

3.6.5. Centrality Measurement

Centrality is "a property of a node's position in a network. We might regard centrality as the structural importance of a node. "(Borgatti, Everett, & Johnson, 2013, p. 164) Among all the measures in social network analysis (SNA), centrality has commonly been applied to different areas such as Sociology, Education, Management, and so on. Further, Kane et al. (2012) emphasize that SNA provides the information to understand the structural features of users in social media in terms of their personal network positions (e.g., Degree Centrality) and the features of the overall network (e.g., Network Density) . Consequently, researchers adopt SNA and use centrality measure to identify influencers.

3.6.6. Tag/Topic/Interest –oriented Algorithms

In a social media network, users participate in discussions based on the topics they have a join interest. Thus, another group of scholars investigates how these topics/interests can be employed for influencer identification. Some social media offer a "tag" function, which allows users to emphasize the topic/interest of their posts. These tags are also recognized as one kind of topic. Zhou et al. (2014) argue that the semantic information in topic-specific content is critical for influencer identification. They focus on a user-network features and the sentiment of text

content to study the online network in the Bulletin Board System (BBS), a traditional form of online forums, for influencer identification. This group of influencer identification research mainly focuses on identifying influential nodes based on the sentiment and/or topic of the text content in their posts.

3.7. Experiment Design

Having reviewed previous approaches to influencer identification, this study designed a set of experiments to compare differences among algorithms/methods. The relative assessment of these approaches provides a roadmap for academicians to move forward on this topic and helps practitioners apply this SMA implementation for influencer identification to real-world situations. This chapter discusses the experiment procedures and results in the next few sections.

3.7.1. Data Collection and Description

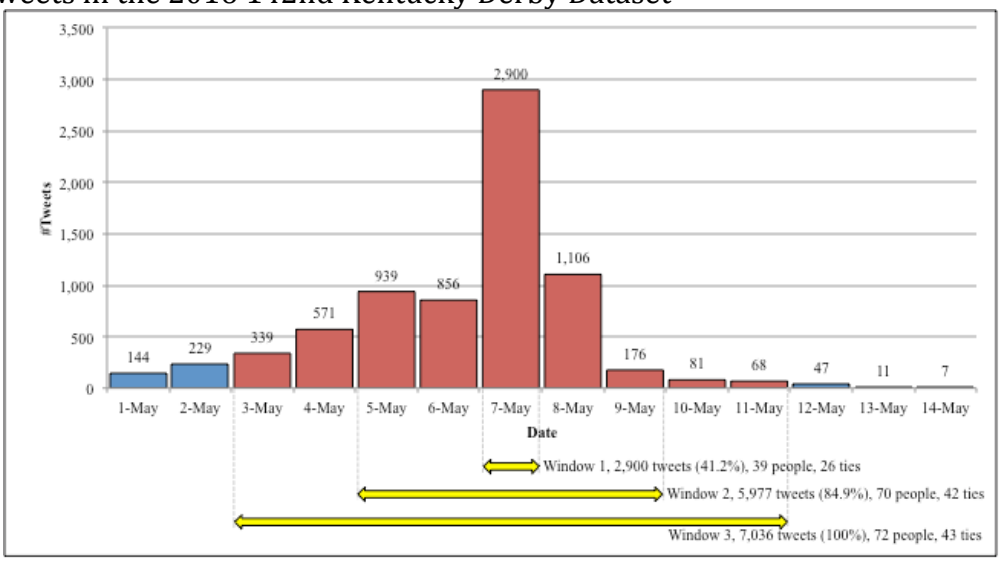
The data for the experiments were gathered by crawling Twitter. The main reasons this dissertation selects Twitter as the data source because (i) it includes relatively sufficient data in text content, social network, and user activity, (ii) its API allows user to efficiently extract data based on keywords, and (iii) its leading position in social media provides a representative result. The Twitter web crawler, which was coded in Python (v2.7.11), was deployed to work in conjunction with the official Twitter Streaming API to extract relevant tweets from Twitter. To do this, the crawler utilizes specific, analyst-supplied keywords to guide its search. This study carefully chose keywords to identify users involved in discussion of a given event of interest over a specified time period.

This study focused on three events: the 2016 U.S. presidential primaries held on March 16, 2016 (with keywords *Hillary*, *Clinton*, *Donald*, and *Trump*), the 2016 March Madness NCAA Basketball Tournaments held between March 21 and March 29, 2016 (with keywords *March* and *Madness*), and the 142nd Kentucky Derby held on May 7, 2016 (with keywords *Kentucky* and *Derby*). In each case, the users are interacting with each other by commenting, replying, or retweeting. Such activities result in the formation of large social media networks in Twitter. The primary purpose of each approach examined here is to help identify the social influencer(s) in each network.

Data were gathered from each source over three time windows containing the event of interest. When dealing with big data, there often is no optimal choice regarding how much data is enough. This study started by gathering data over an extended period encompassing several days around the event of interest and plotting the frequency of tweets on each day within this time span. This study then created three data sets, labeled *Window 1*, *Window 2*, and *Window 3*. Window 1, the smallest window, focuses on all tweets gathered on the day with the largest number of tweets, Window 2 focuses on all tweets in the 5-day period centered on Window 1, and Window 3 focuses on all tweets in the 9-day period centered on Window 1. Thus, for each source, this study created data sets of 3 sizes as shown in the Figure 3.4:

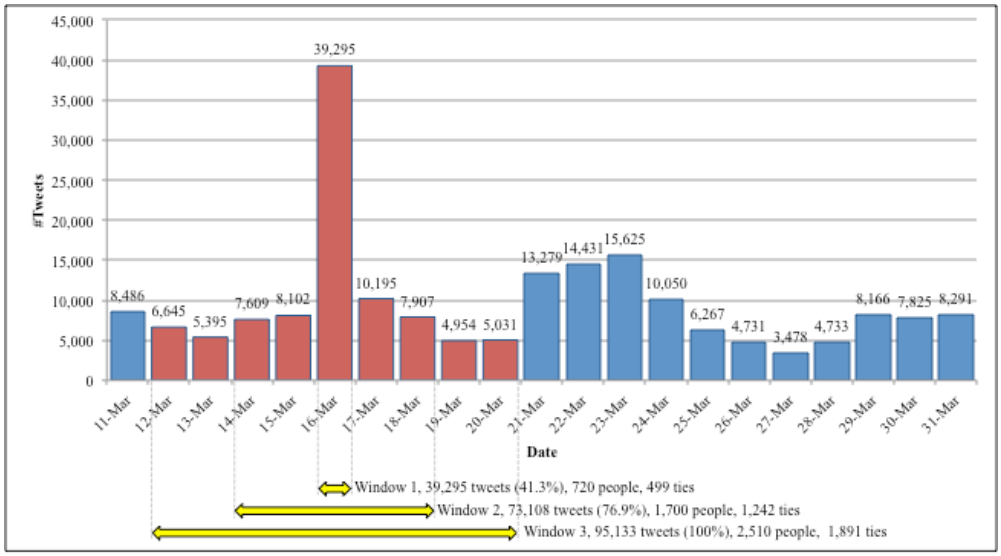
Figure 3.4 Data Window Constructions

(a) Tweets in the 2016 142nd Kentucky Derby Dataset



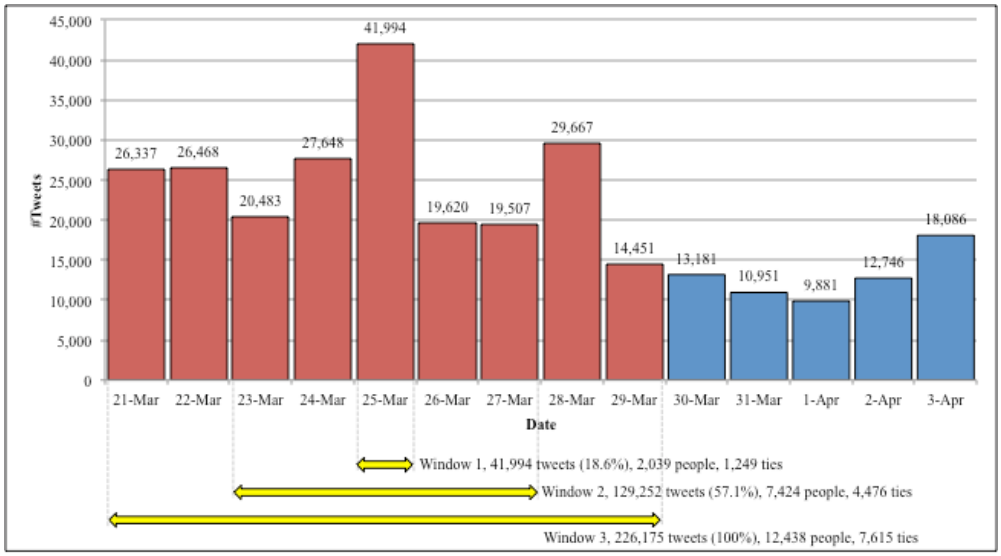
*The 142nd Kentucky Derby held on May 6 & May 7, 2016 (keywords: Kentucky and Derby)

(b) Tweets in the 2016 U.S. Presidential Primaries Dataset



*The 2016 US Presidential Primaries held on March 15, 2016 (keywords: Hillary, Clinton, Donald, and Trump)

(c) Tweets in the 2016 March Madness NCAA Basketball Tournaments Dataset



*The 2016 March Madness NCAA Basketball Tournaments held between March 17 and April 4, 2016 (keywords: March and Madness)

The data was captured in the JSON (JavaScript Object Notation) format and subsequently stored in a MongoDB (v2.6.12) database for further analysis. MongoDB is used because it allows importing and storing collected data without the need for remapping (Kumar et al., 2014). The experiments for evaluating each algorithm in this study were processed on Microsoft Azure Virtual Machines (Standard_D1) running the Linux OS (Ubuntu server 12.04.5 LTS). This allowed me to run multiple experiments independently and in parallel under identical experimental conditions. Because of the limitation of the virtual machines, it may exist slight difference in the computation times between these virtual machines. Hence, the computation

times measured in this dissertation can only provide similar results to the real-world situation. However, the outcomes from multiple datasets show the consistent results, and this can support that the slight differences between different virtual machines will not affect the conclusions provided in this dissertation.

3.7.2. Evaluation Metrics

The main goal of this study was to compare different influencer identification approaches in the same setting in order to examine the performance of each approach. This study evaluated the performance using two main criteria: computation time and the quality of the identified influencers. The main argument for these criteria is that an approach that identifies social media influencers more quickly and provides a better quality of influencers is a better-performing approach.

3.7.2.1 Computation Time

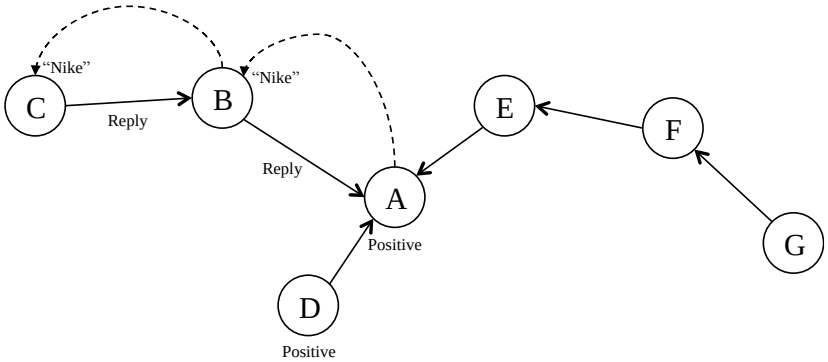
To record the computation time for each experiment, this study uploaded experimental data from the MongoDB storage to each virtual machine using an FTP server. Each virtual run stored information about the run start and end times, run execution time, and I/O time. The results per run provide information for understanding each node's ranking. In each run, the run execution time and generated iterative outputs are captured until the algorithm reached a state of convergence. The benefit of doing this was that it enabled me to determine whether the top ranked influencers were identified in the early iterations. If the goal is only to identify the top *N* percent of social influencers (e.g., top 5%, 10%, or 20% of the influencers), it might be not necessary to run each approach until it reaches a state of convergence. The run execution time per iteration, when compared to the total execution time, provides information about how much time can be saved if the approach stops during the early iterations.

3.7.2.2 The Quality of Identified Influencers

To date, there is not a standard metric to measure the quality of influencers identified by these algorithms/methods. After reviewing the literature, this dissertation documented detailed information about the different identification approaches and the metrics used in these papers to evaluate the quality of the identified influencers in Appendix II. However, some metrics used in these papers are not applicable in this study. For example, precision, recall, and F-measure methods are used in some research. Precision measures the possibility that a classifier labels as

positive a node that is actually negative, and recall measures the ability of a classifier to identify all of the positive nodes. The F-measure is the weighted harmonic mean of the precision and recall. These three metrics are only applicable when there is a correct answer on which the metrics can be based, and these papers use artificial judgments for the basis. This strategy is not available when a dataset is relatively large. Hence, this study adopts multiple applicable metrics from the literature to evaluate the results of different identification approaches. These metrics provide alternative views for explaining the advantage of each approach. The explanations of these metrics follow:

Figure 3.5 Illustration for Understanding Evaluation Metrics



1. Coverage rate: The coverage of a node is the number of nodes directly or indirectly connected to it. In the example in Figure 3.5, node A directly or indirectly covers nodes B, C, D, E, F, and G because they all directly (B, D, and E) or indirectly (C, F, and G) connect to node A (i.e., coverage = 6). Node E in turn covers nodes F and G (i.e., coverage = 2). The coverage rate for a node is the ratio of the node's coverage value to the sum of the coverage values of all nodes in the network.
2. Language diffusion rate: If node A included the term *Nike* in a tweet and node B also said *Nike* in his/her reply, the tweet from A is considered to be influential and the information has been diffused to B. If node C, in replying to node B, also includes *Nike*, then node C can be considered to be influenced by node A through node B. The total numbers of words that are repeated in subsequent replies are calculated for each individual posted tweet. Each node may have multiple posts. The number of repeated words appearing in subsequent replies to each post is summed for each node. The language diffusion rate for a node is the ratio of the node's language diffusion value to the sum of the language diffusion values of all nodes in the network.

3. Agreement rate: Sentiment analysis categorizes tweets into three types: positive, negative, or neutral. If node D replies to node A's tweet with positive sentiment, this indicates that node D agrees with node A and it counts in the agreement value of A. The value of a node's agreement is the sum of the replies that agree with the node. Each node's agreement rate is the ratio of the node's agreement value to the sum of the agreement values of all nodes in the network.

3.7.3. Experiment Procedure

Using the metrics just described, this study analyzed three datasets in three different data windows (3×3 datasets). The main purpose of executing the influencer identification approaches to multiple datasets was to check the generalizability of the experiment results. All of the identification approaches that were examined in this study were coded in Python and executed on multiple virtual machines with the same setting in order to obtain experiment results and record computation times based on the exact same computer setting. Each approach was applied to analyze each dataset in order to identify the influencers and their rankings. The ranking of each influencer was then used to extract the top *N*% of the identified influencers.

This study examines fifteen approaches to social influencer identification from four main categories: HITS-based algorithms (HT1and HT2), PageRank-based algorithms (PR1, PR2, PR3, PR5, PR6, PR7, PR8, and PR9), clustering-based algorithms (CL1, CL2, and CL3), and centrality-based approaches (CE1 and CE2). Among the PageRank-based algorithms, algorithm PR4 was excluded from the experiment because the main data this algorithm uses—a "trust" score to evaluate each user and identify the influencers in the portal—were absent in the datasets of this study. The trust score is measured by the website used in the study that proposed PR4 (Eopinion.com), and Twitter does not support such a score. Another algorithm, PR9, had not reached the status of convergence at the point when the experiment results were reported. It is possible that this algorithm could not reach convergence and identify the influencers in a huge social media network because of the machine setting in this study. Further experiments should be done with a more advanced computer to obtain results for algorithm PR9.

After executing each approach to influencer identification using each window, this study measured the performance of the approaches based on (1) computation time and (2) the three previously described metrics. The computation times show differences in the performance of the approaches when identifying influencers in the same dataset, and the quality metrics indicate the

quality of the influencers identified by each algorithm/method. The following section discusses the results.

3.8. Discussion

This study gathered results for three different windows in three different datasets. Here, this discussion is results for the Election dataset from Twitter, and this dissertation presents the results for the other two datasets in Appendices III and IV. Because of the medium size of the Twitter Election dataset, these results provide an overview of the approaches to influencer identification in terms of their different performances with respect to computation time and the quality of the influencers they identify. The capital *N* showing in all the results represents the population of influencers in each data set and the lower case *n* refers to the number of influencers selected (e.g., top 1% of influencers in the Twitter election dataset includes 7 people).

3.8.1. Results: Computation Time

Using the experiment design described above, this study recorded computation times for each approach using the same machine setting and datasets. The purpose was to understand the exact moment when the top *N*% of the influencers have been identified. As noted in earlier section, a given approach may not be required to run to completion to identify the top *N*% of the influencers, which mainly depends on specific requirements. For example, if an approach actually finds the top 1% of the influencers (i.e., the top 7 influencers in window 1 of the Twitter Election dataset) in an early iteration (e.g., the third iteration), this study captures the time it takes to identify those top 7 of influencers. The following discussions present the results for the Twitter Election dataset to illustrate the computation time differences between the different approaches.

Table 3.3 Computation Time for Finding the Top *N*% of Influencers: Twitter Election Dataset

Window 1, *N*=720 (Unit: Minutes)					
	1% (*n*=7)	**5% (*n*=36)**	**10% (*n*=72)**	**20% (*n*=144)**	**100% (*N*=720)**
HT1	275.12	275.12	275.12	275.12	275.12
HT2	318.96	318.96	318.96	318.96	318.96
PR1	77.00	77.00	77.00	77.00	77.00
PR2	8.99	13.41	13.41	13.41	13.41
PR3	59.58	59.58	59.58	59.58	59.58
PR5	13.43	17.78	17.78	17.78	17.78

PR6	0.38	0.39	0.39	0.39	0.39
PR7	10.26	10.26	19.89	19.89	19.89
PR8	3.42	3.42	3.42	3.42	10.03
CE1	0.30	0.30	0.30	0.30	0.30
CE2	0.30	0.30	0.30	0.30	0.30
CL1	0.96	0.96	0.96	0.96	0.96
CL2	2.88	2.88	2.88	2.88	2.88
CL3	3.00	3.00	3.00	3.00	3.00
Window 2, *N*=1,700 (Unit: Minutes)					
	1% (*n*=17)	**5% (*n*=85)**	**10% (*n*=170)**	**20% (*n*=340)**	**100% (*n*=1700)**
HT1	862.82	862.82	862.82	862.82	862.82
HT2	974.30	974.31	974.31	974.31	974.31
PR1	368.04	368.04	368.04	368.04	368.04
PR2	35.69	76.98	76.98	76.98	76.98
PR3	1185.38	1185.38	1185.38	1185.38	1185.38
PR5	69.75	132.74	132.74	132.74	132.74
PR6	2.33	2.33	2.33	2.33	2.33
PR7	43.33	64.48	64.48	64.48	64.48
PR8	17.37	17.37	17.37	17.37	50.50
CE1	0.77	0.77	0.77	0.77	0.77
CE2	0.77	0.77	0.77	0.77	0.77
CL1	4.20	4.20	4.20	4.20	4.20
CL2	12.56	12.56	12.56	12.56	12.56
CL3	16.66	16.66	16.66	16.66	16.66
Window 3, *N*=2,510 (Unit: Minutes)					
	1% (*n*=25)	**5% (*n*=125)**	**10% (*n*=251)**	**20% (*n*=502)**	**100% (*n*=2,510)**
HT1	2278.92	2278.92	2278.92	2278.92	2278.92
HT2	1586.42	1586.42	1586.42	1586.42	1586.42
PR1	758.20	758.20	758.20	758.20	758.20
PR2	108.84	150.00	150.00	150.00	150.00
PR3	1569.15	1569.15	1569.15	1569.15	1569.15
PR5	167.40	209.12	209.12	209.12	209.12
PR6	10.17	10.17	10.17	10.17	10.17
PR7	123.26	123.26	123.26	123.26	123.26
PR8	167.85	167.85	167.85	167.85	505.92
CE1	3.19	3.19	3.19	3.19	3.19
CE2	3.19	3.19	3.19	3.19	3.19
CL1	8.24	8.24	8.24	8.24	8.24
CL2	23.70	23.70	23.70	23.70	23.70
CL3	32.56	32.56	32.56	32.56	32.56

* All the source code can be downloaded from following link:
https://www.dropbox.com/s/cd8c87ijat3u55n/source_code.docx?dl=0

Table 3.3 provides a simple idea of the different performances of these approaches in terms of computation tine. The entries in the table show how many minutes it took each approach to find influencers and reach a state of convergence. The computation times for the fastest and the slowest approaches turn out to be dramatically different. For example, when identifying the top 20% of the influencers in Window 3, the slowest approach (HT1, 2278.92 minutes) takes over 700 times longer than the fastest two approaches (CE1 and CE2, 3.19 minutes). In addition, most of the algorithms/methods could not find the top N% of influencers in early iterations. The exceptions were (1) PR2, PR5, and PR6 for Window 1; (2) PR2, PR5, and PR7 for Window 2; and (3) PR2 and PR5 for Window 3. The table shows that HT1, HT2, and PR3 took longer than the other approaches to identify the top 1%, 5%, 10%, and 20% across the three windows, while CE1 and CE2 took the shortest time across all three windows. PR1 took a long time to identify the top 1%, 5%, 10%, and 20% influencers in window 1, which represents that PR1 preforms better than other approaches only in bigger network. Generally speaking, HITS-based algorithms take much longer for influencer identification, while centrality-based methods are relatively time efficient. However, time is only one measurement when considering performance. In the next section, this study discusses results related to the quality of the identified influencers in order to compare the goodness of results from alternative approaches.

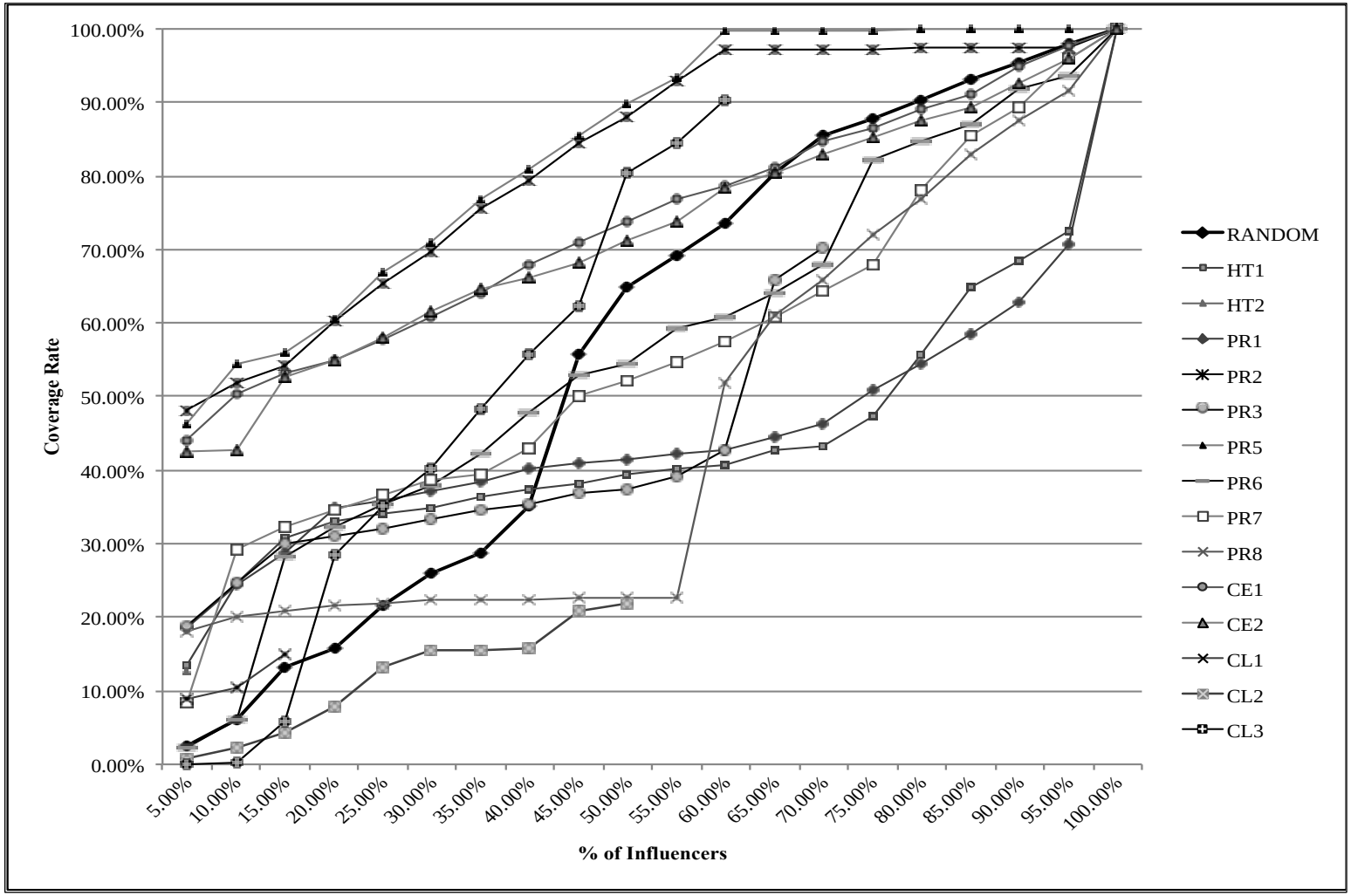
3.8.2. Results: The Quality of Identified Influencers

3.8.2.1 Coverage Rate

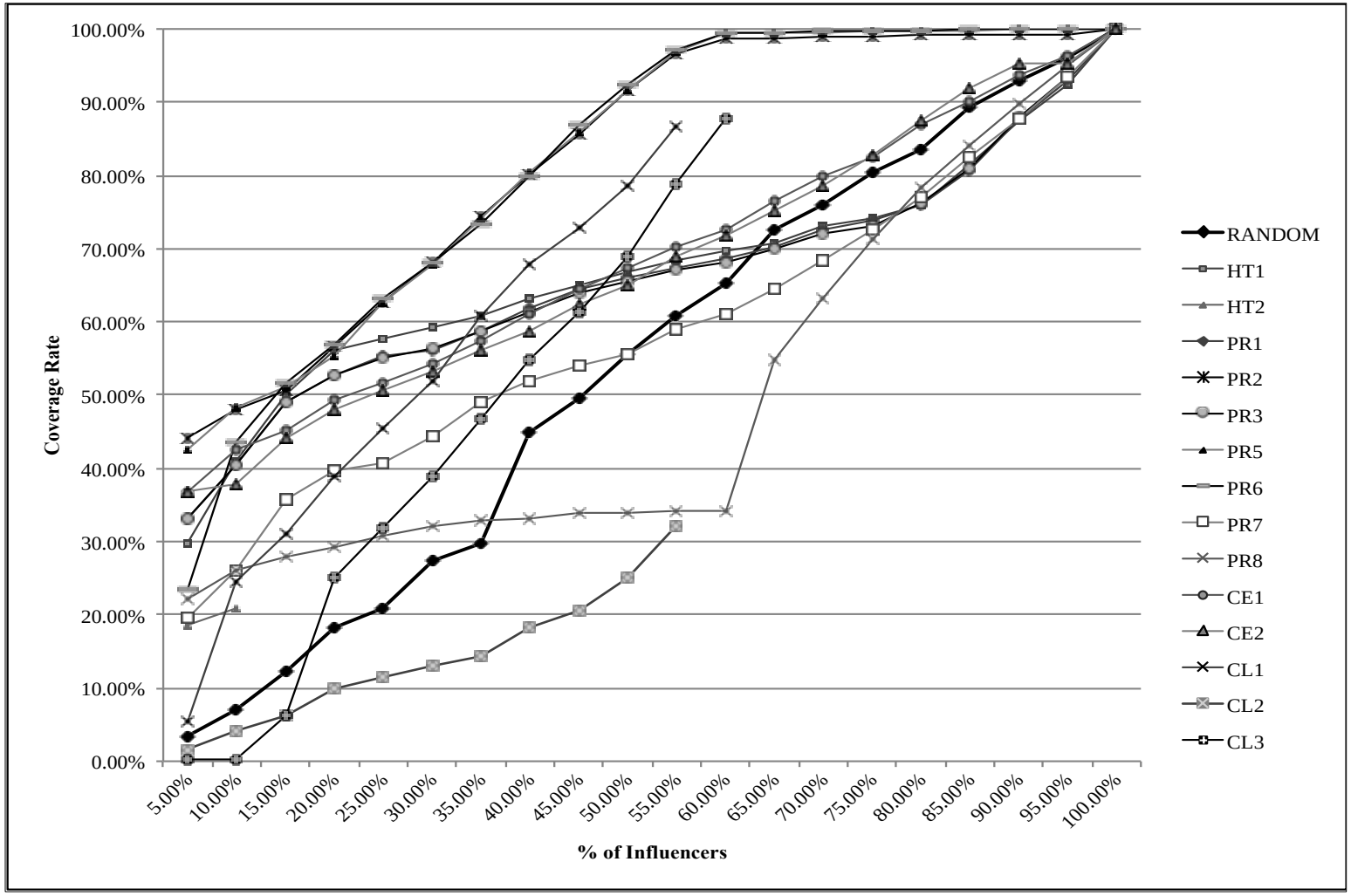
The coverage rate indicates the percentage of users covered by the top *N*% of influencers identified by each algorithm/method. The line charts shown in Figure 3.6 depict the overall results for the identified influencers' coverage rates for the three data windows. Each line chart includes all approaches to show the overall differences.

Figure 3.6 Coverage Rate for Different Algorithms: Twitter Election Dataset

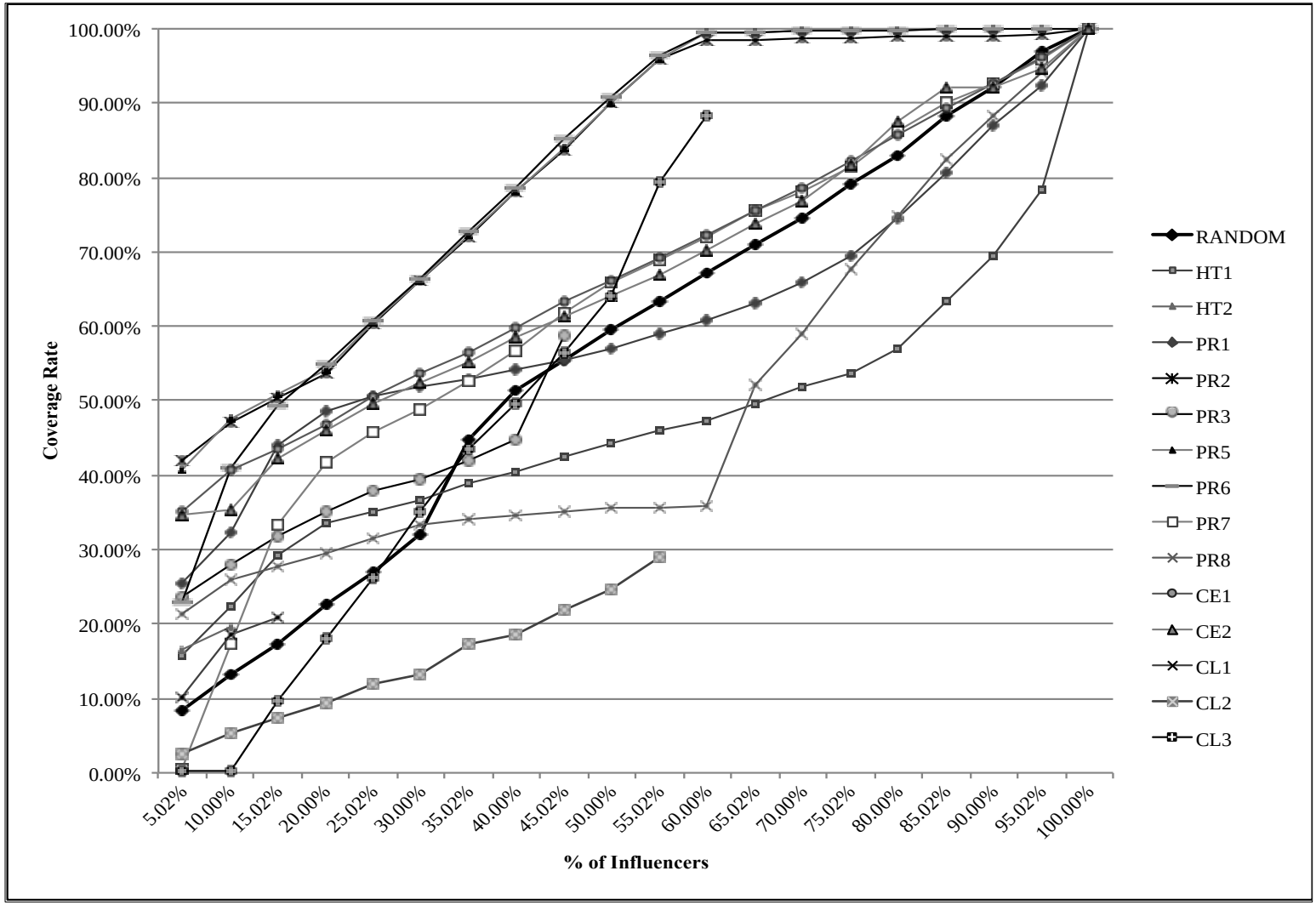
(a) Window 1, *N*=720



(b) Window 2, *N*=1,700



(c) Window 3, *N*=2,510



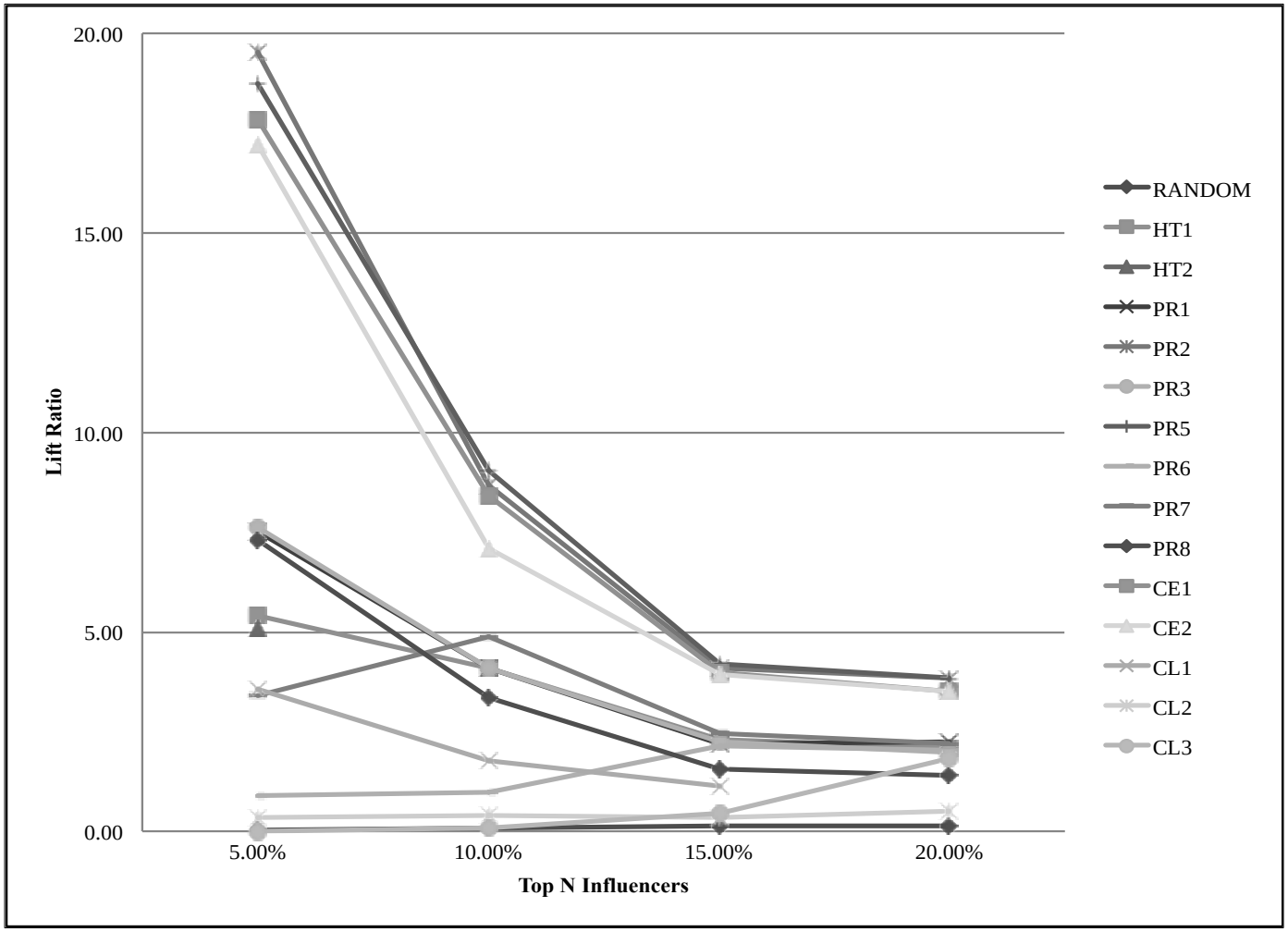
The line charts in Figure 3.6 use the x-axis for the cumulative percentage of influencers and the y-axis for the cumulative coverage rate to show the correlation between these. For example, in Window 3, the top 20% of influencers identified by CL2 represent around a 10% coverage rate, which means that the information from these 20% of the influencers will reach around 10% of other users. The trend of the coverage rate by influencers shows that when the proportion of influencers increases, the number of people these influencers cover also increases. A random sampling method is also executed in each experiment to present a random line for comparison.

Overall, most of the approaches are size sensitive, which means that when the data size changes, the resultant coverage rate changes correspondingly. Most approaches perform better than the random sampling method in terms of the coverage rate when the influencer percentage is small, with the exception of algorithm PR8 and CL2. The random line shows a relatively steep slope across the three different windows.

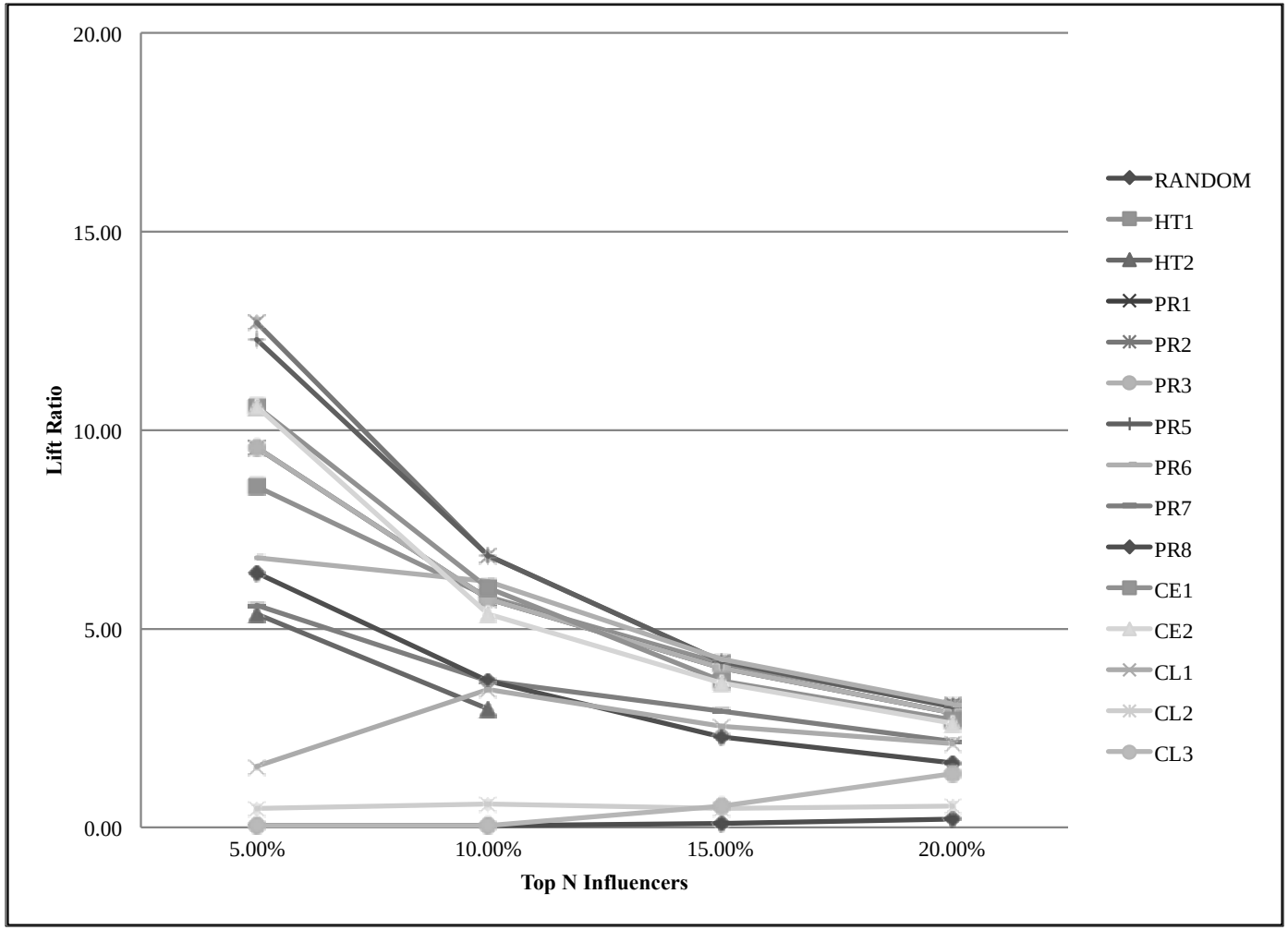
The line charts in Figure 3.6 present the trend of each approach from 0% to 100% of the influencers. However, it is more practical to identify the top 20% of influencers because the social media network is huge. Figure 3.7 presents the top 20% of the influencers in lift ratio charts to show the differences between how these approaches perform. The lift ratio charts record the ratio of differing coverage rates of each approach and the random sampling approach.

Figure 3.7 Coverage Rate Lift Ratio Charts for Different Algorithms: Twitter Election Dataset

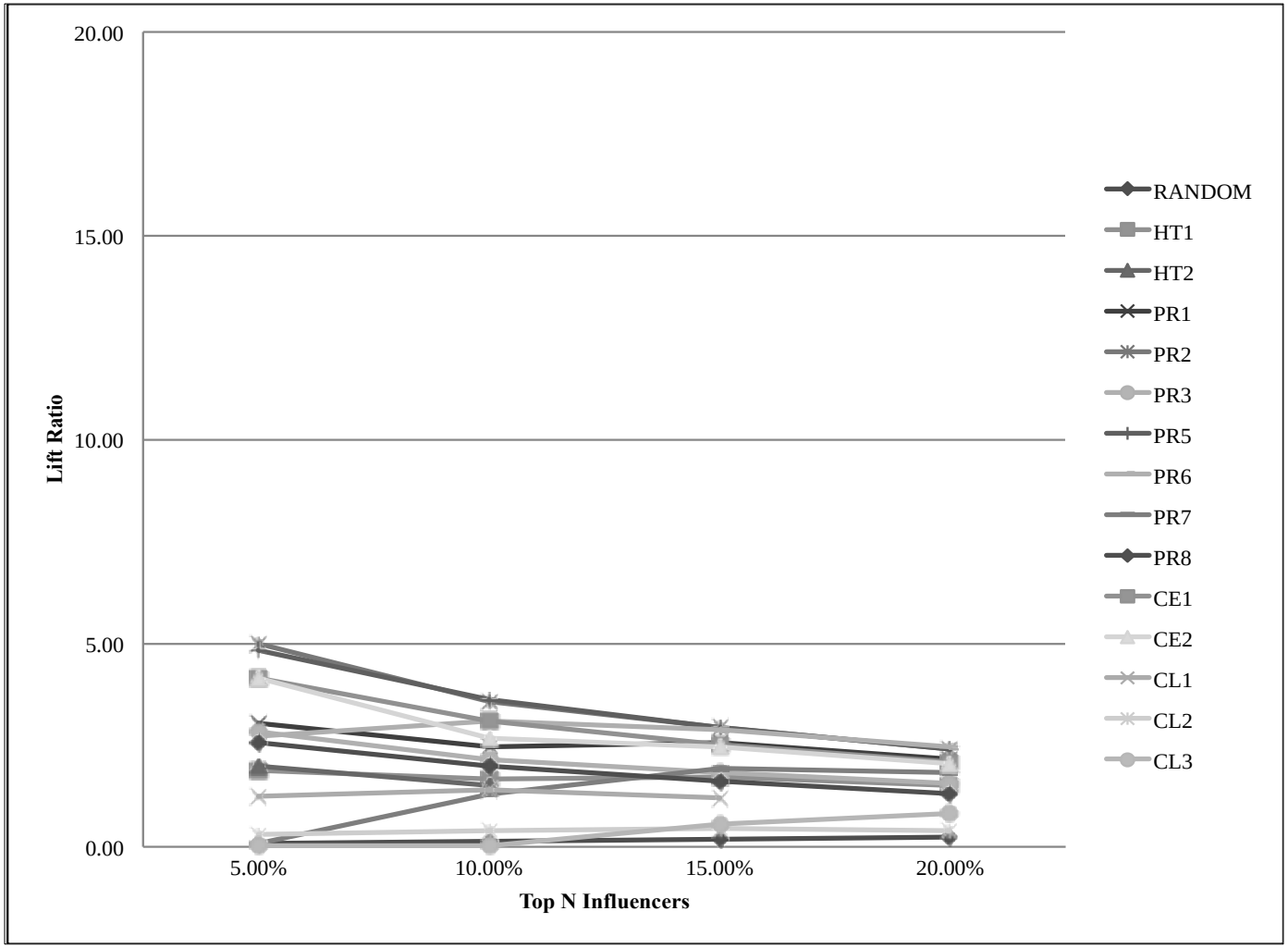
(a) Window 1, *N*=720



(b) Window 2, *N*=1,700



(c) Window 3, *N*=2,510



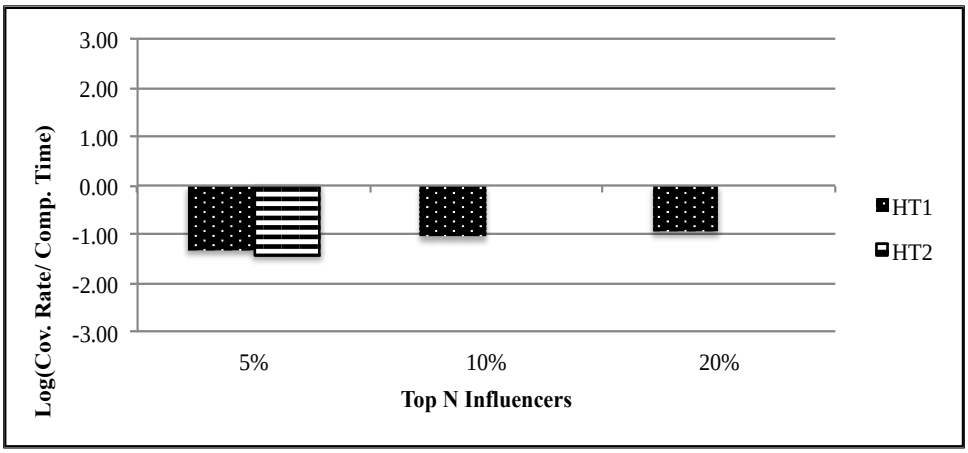
The lift ratio charts in Figure 3.7 show that the top 20 % of the influencers identified by PR2, PR5, and PR6 yield a better coverage rate than those identified by the random sampling method. CE1 and CE2 also perform well in identifying the top 20% of influencers in terms of coverage rate. These lift ratio charts provide a relatively close look at the identification of the top 20% of influencers. In order to provide a clearer picture about the performance of each approach, this study also took the time factor into consideration. First captured the computation time for identifying the top 5%, 10%, and 20% of influencers, along with the coverage rate for each approach. Applying the bang-to-buck method to calculate the ratio of the coverage rate to computation time produces the charts in Figure 3.8. These indicate the performance of each approach based on this bang-to-buck ratio.

In these charts, the lift ratio lines reveal a trend of converge when the number of identified influencer increase, which means all the lines are getting closer to each other. This dissertation argues that this represents when identifying a small group of top influencer (e.g., top 5% of influencers), all approaches provide much different quality of influencers in coverage rate. When the number of identified influencer increases, all approaches locate a similar group of influencers. This result provides a view that different approach provides a great difference in identifying small group of influencers, regarding their coverage rate. If the purpose is to identify a relatively large group of influencer, there is another factor can be considered, computation time.

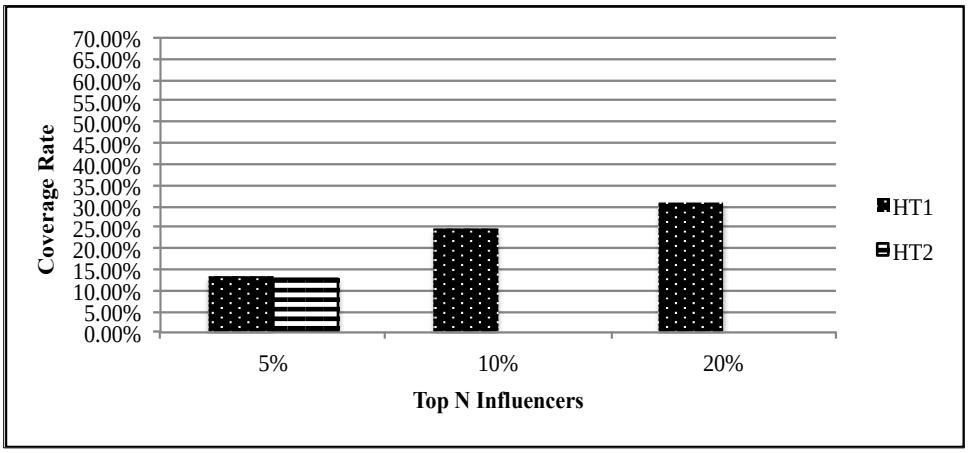
In the bang-to-buck bar charts, this dissertation records the coverage rate and computation time of each approach and put them together based on categories (e.g., PageRank-based algorithms). The bang-to-buck idea is to divide the coverage rate by computation time of each identification approach. Here, this dissertation also applies log transformation due to the huge variation in results for different approaches. The higher bang-to-buck ratio represents the better efficiency in identifying influencers. The results show that most of bang-to-buck ratio are lower than zero.

Figure 3.8 Coverage Rate of Top *N*% of Influencers for Different Algorithms in Different Categories: Twitter Election Dataset

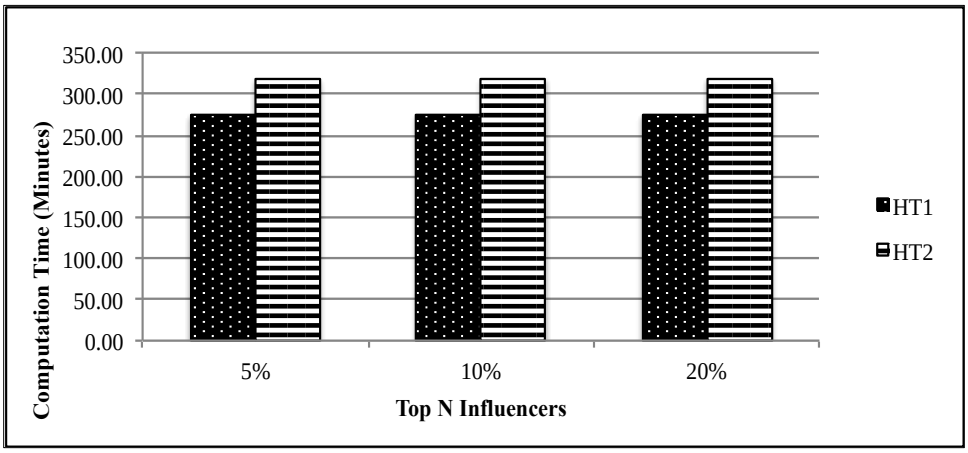
(a-1) HITS-based Algorithms, Window 1, *N*=720



Coverage Rate/ Computation Time

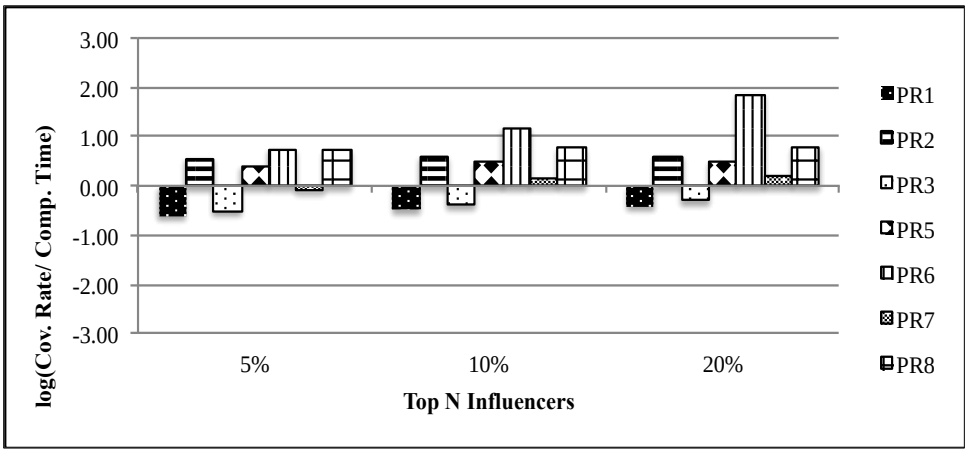


Coverage Rate

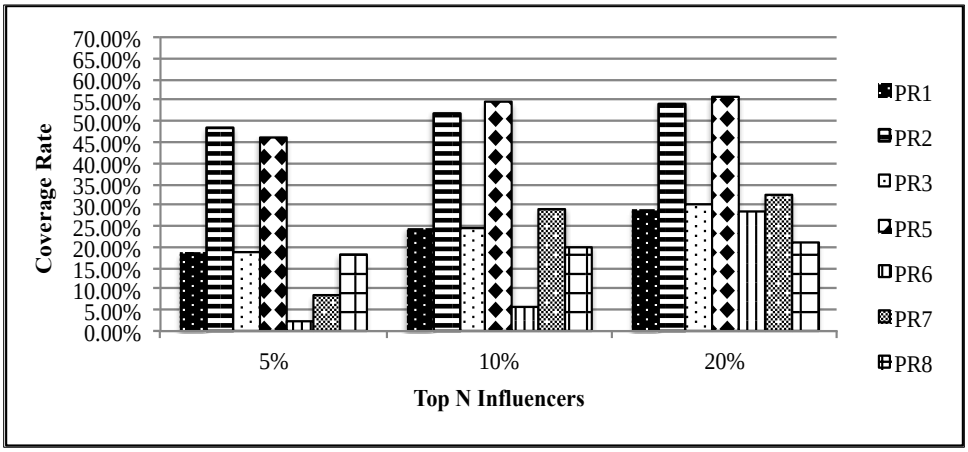


Computation Time

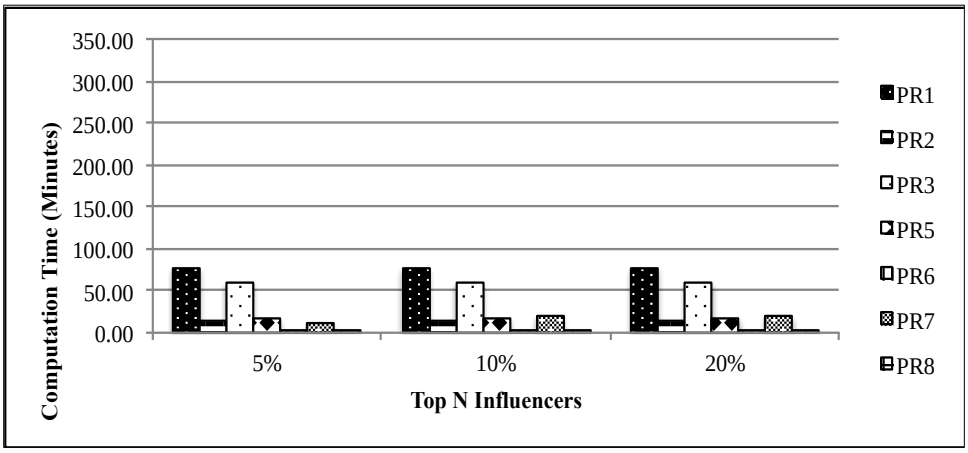
(a-2) PageRank-based Algorithms, Window 1, *N*=720



Coverage Rate/ Computation Time for

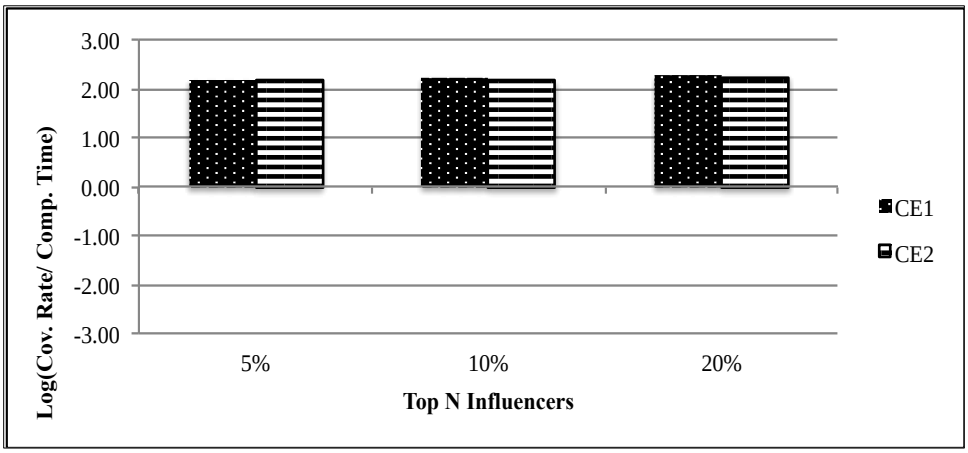


Coverage Rate

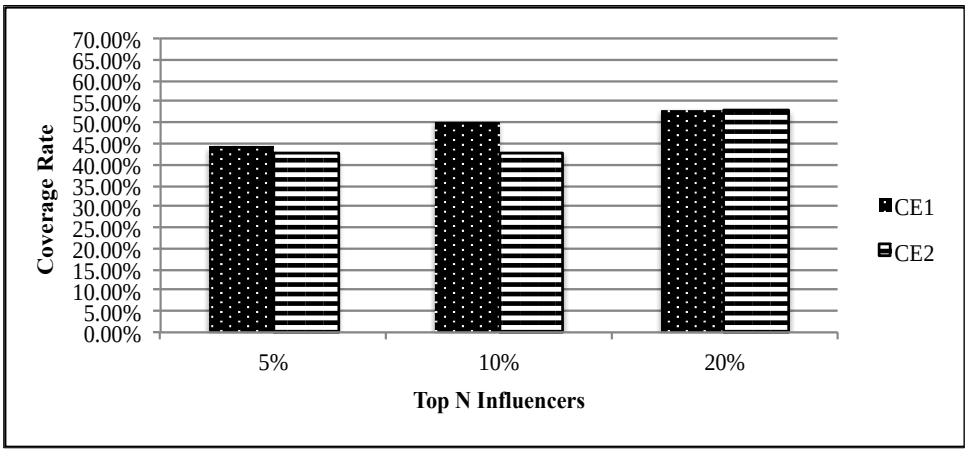


Computation Time

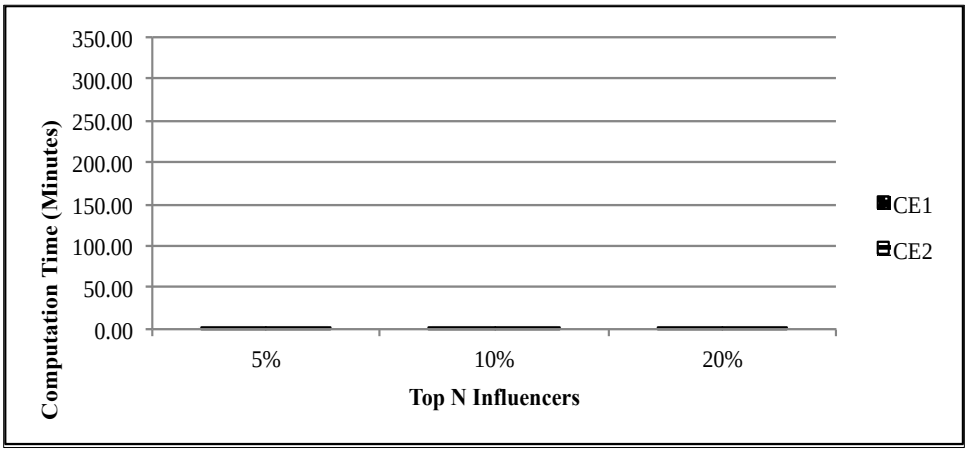
(a-3) Centrality-based Mechanisms, Window 1, *N*=720



Coverage Rate/ Computation Time

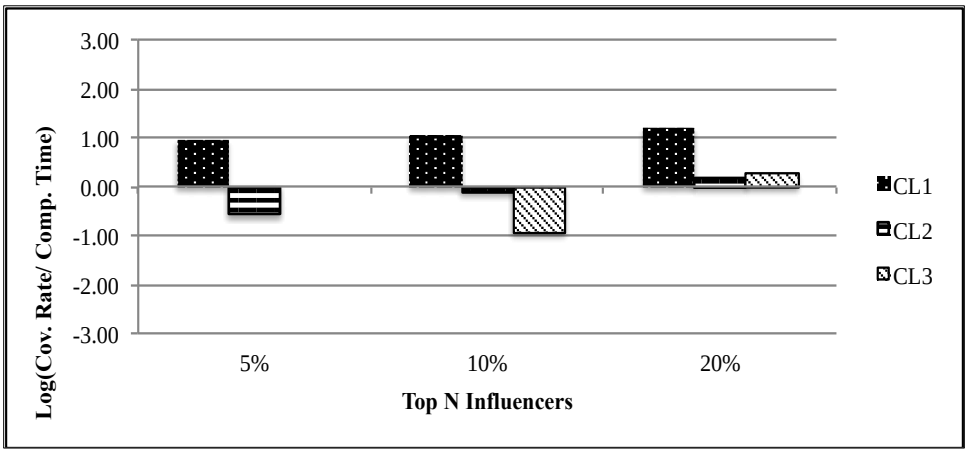


Coverage Rate

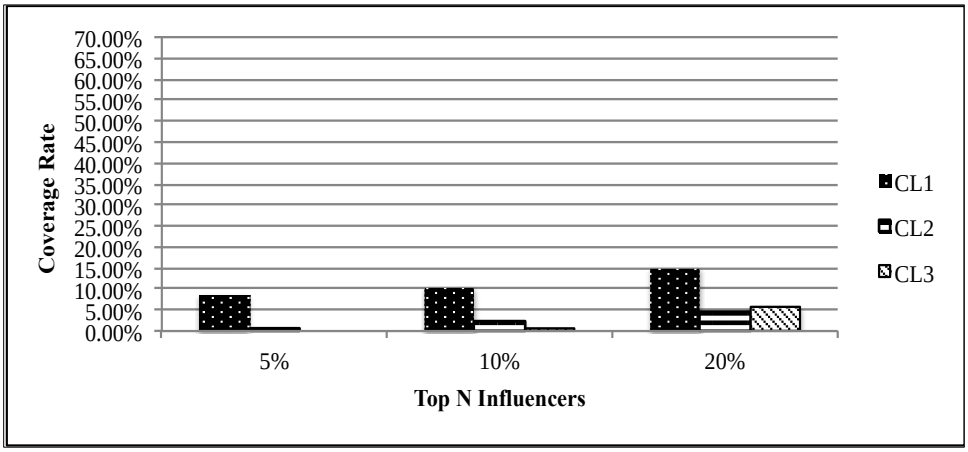


Computation Time

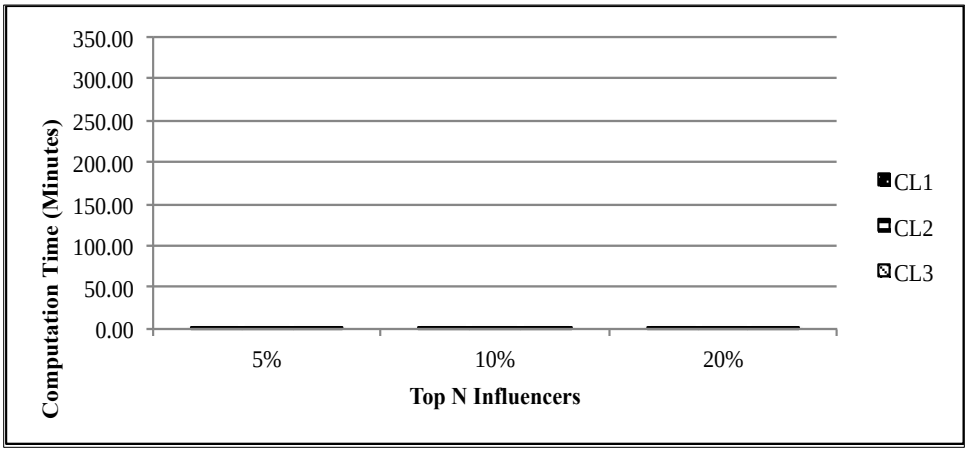
(a-4) Clustering-based Algorithms, Window 1, *N*=720



Coverage Rate/ Computation Time

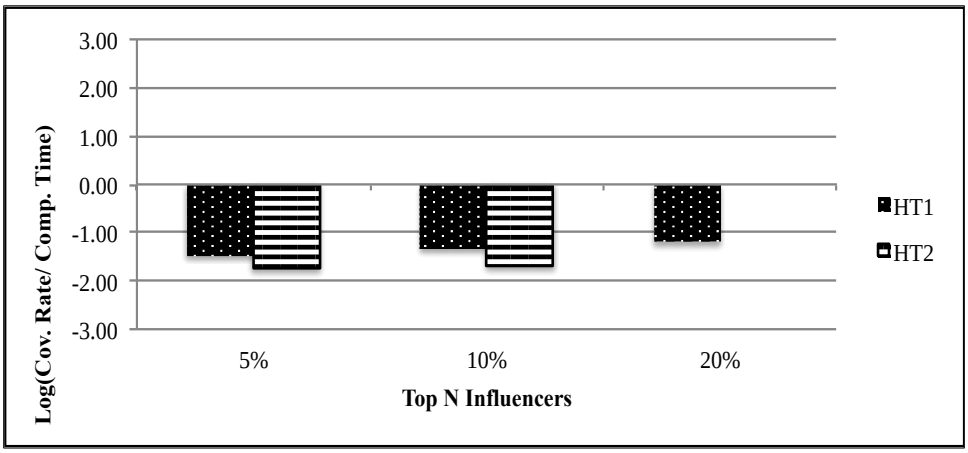


Coverage Rate

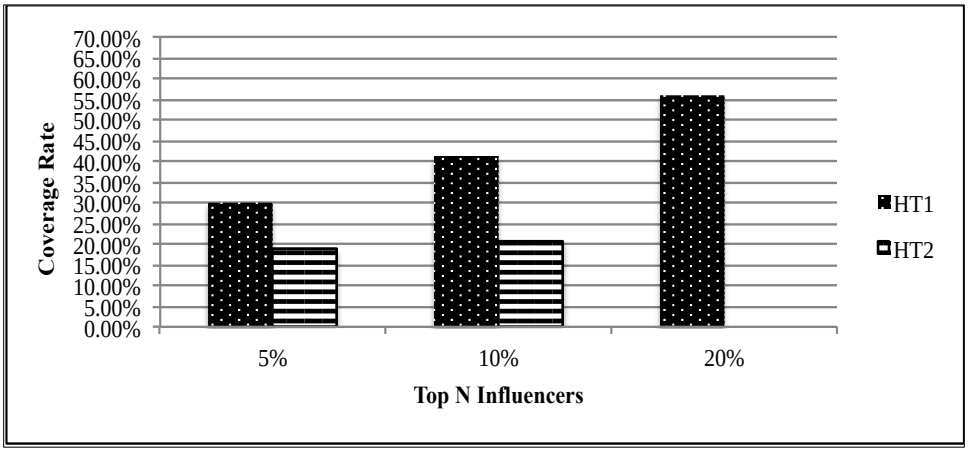


Computation Time

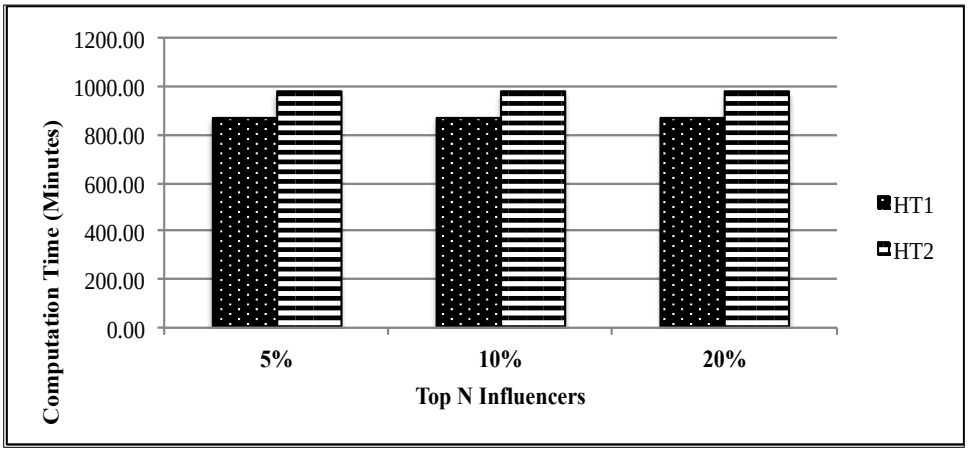
(b-1) HITS-based Algorithms, Window 2, *N*=1,700



Coverage Rate/ Computation Time

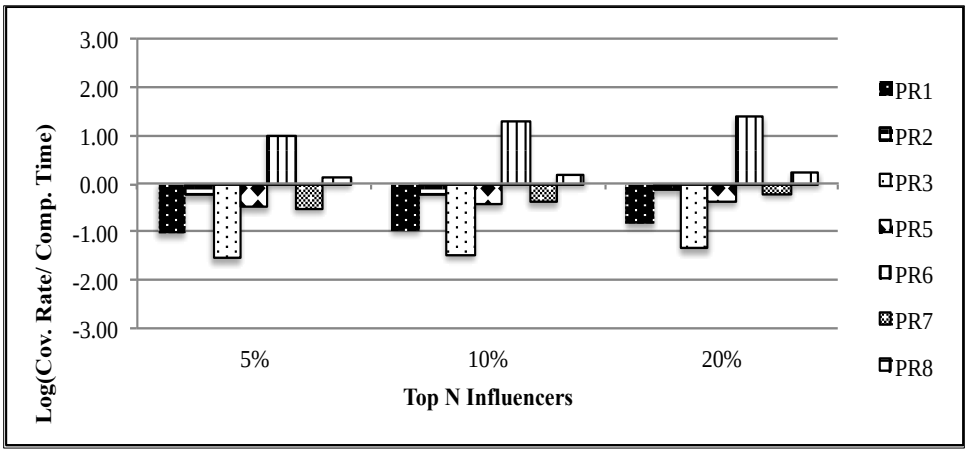


Coverage Rate

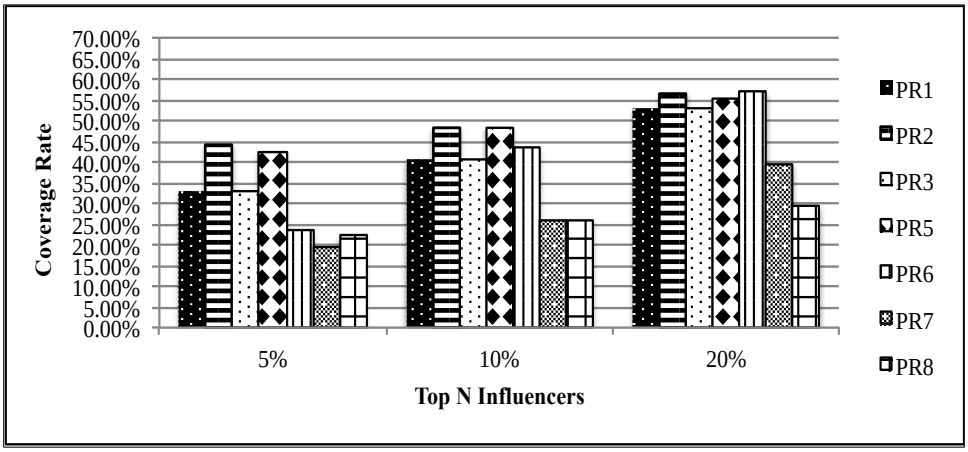


Computation Time

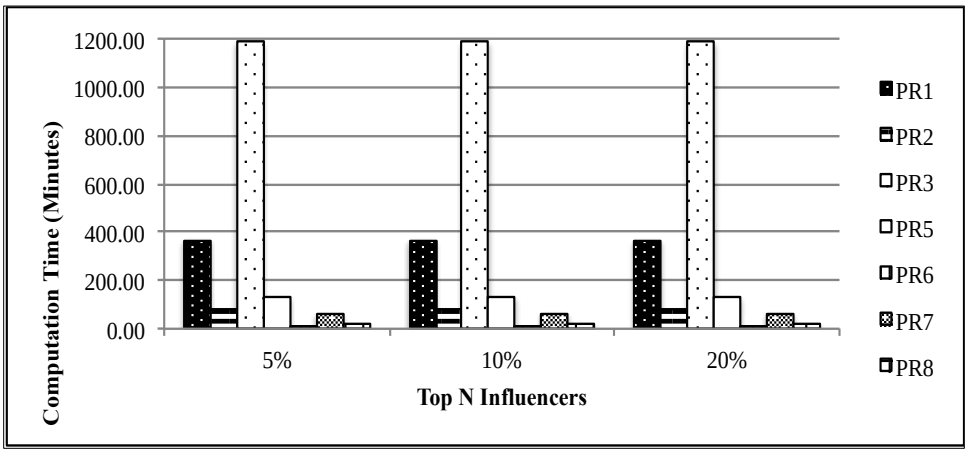
(b-2) PageRank-based Algorithms, Window 2, *N*=1,700



Coverage Rate/ Computation Time

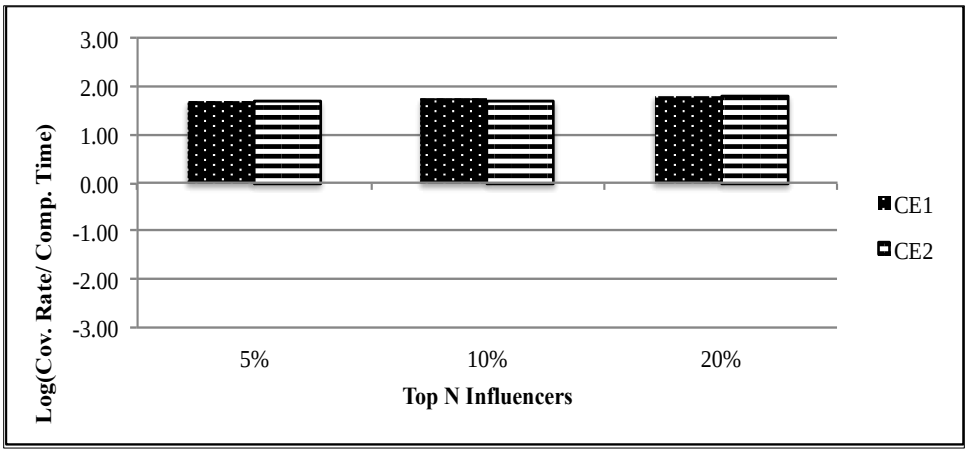


Coverage Rate

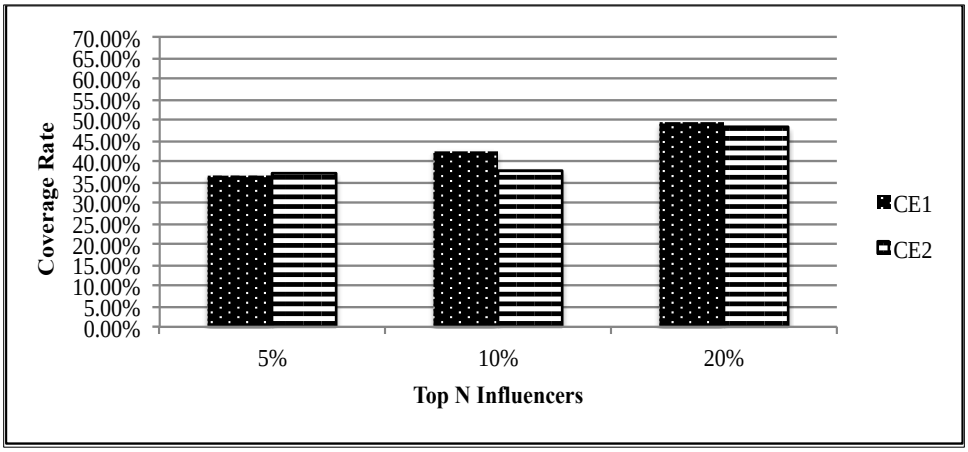


Computation Time

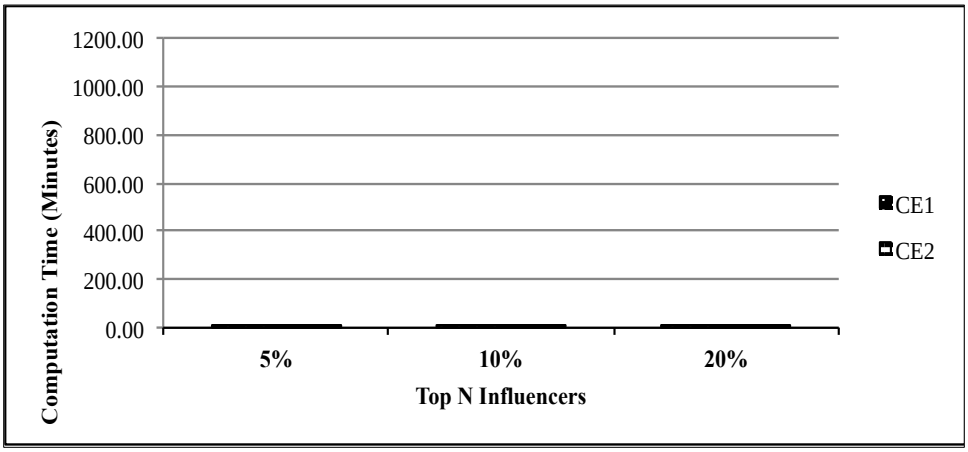
(b-3) Centrality-based Mechanisms, Window 2, *N*=1,700



Coverage Rate/ Computation Time

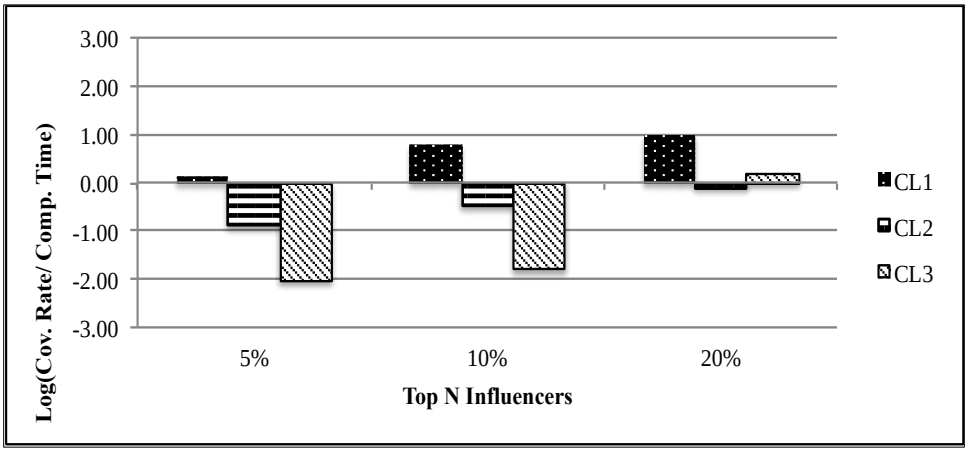


Coverage Rate

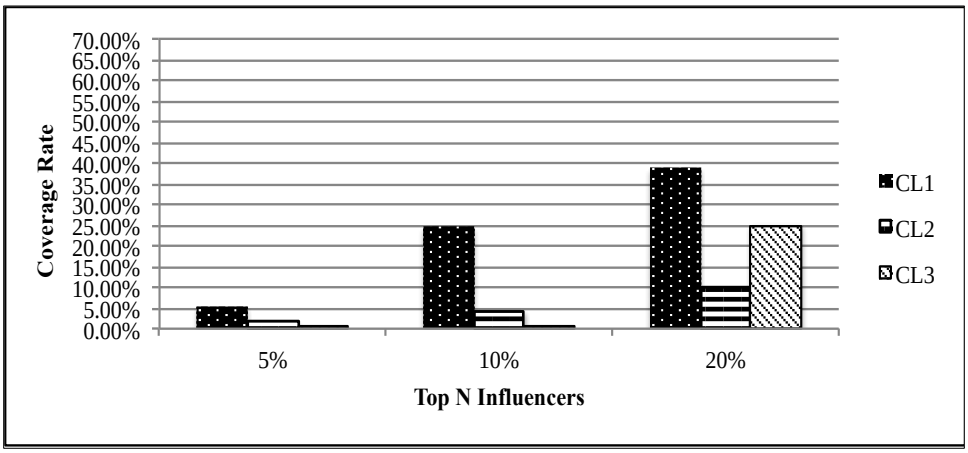


Computation Time

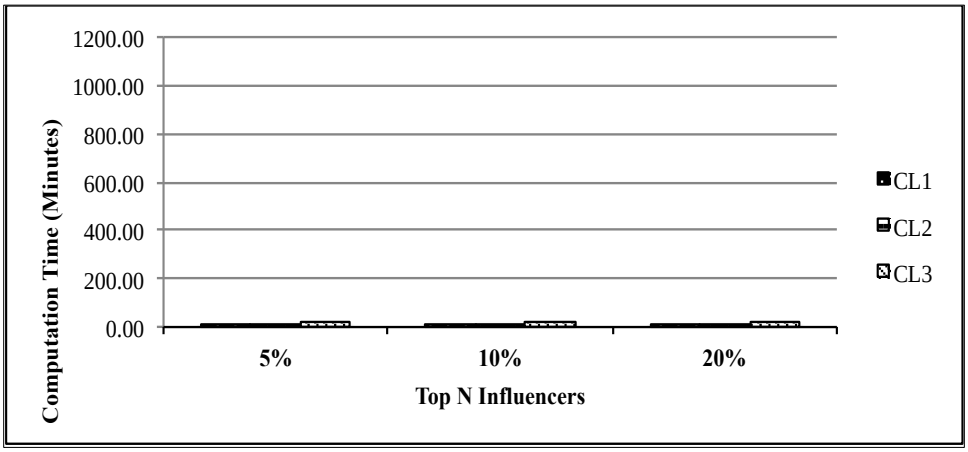
(b-4) Clustering-based Algorithms, Window 2, *N*=1,700



Coverage Rate/ Computation Time

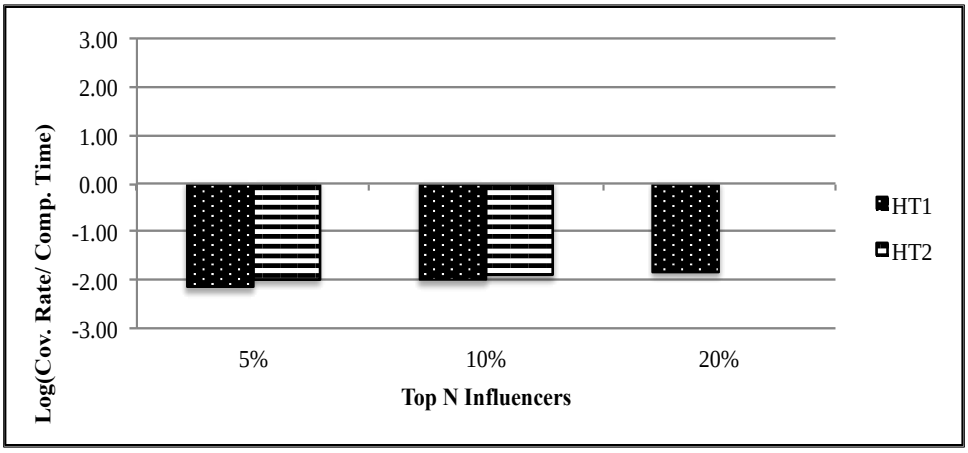


Coverage Rate

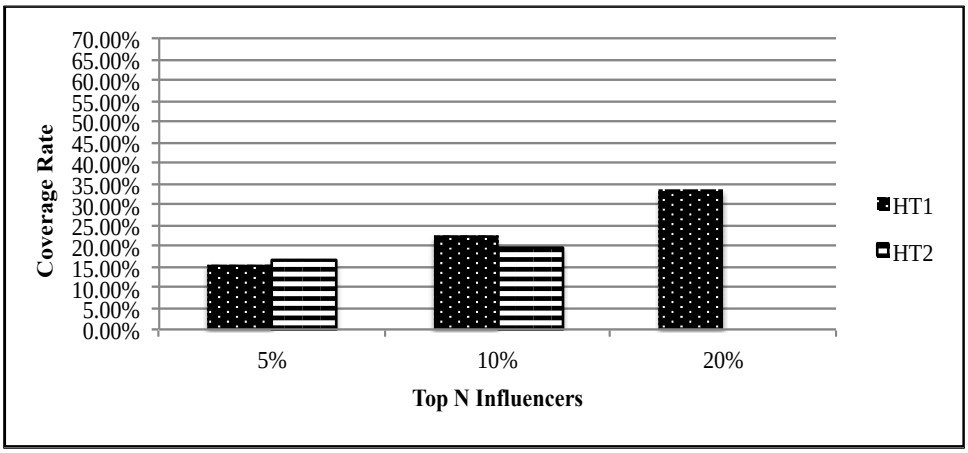


Computation Time

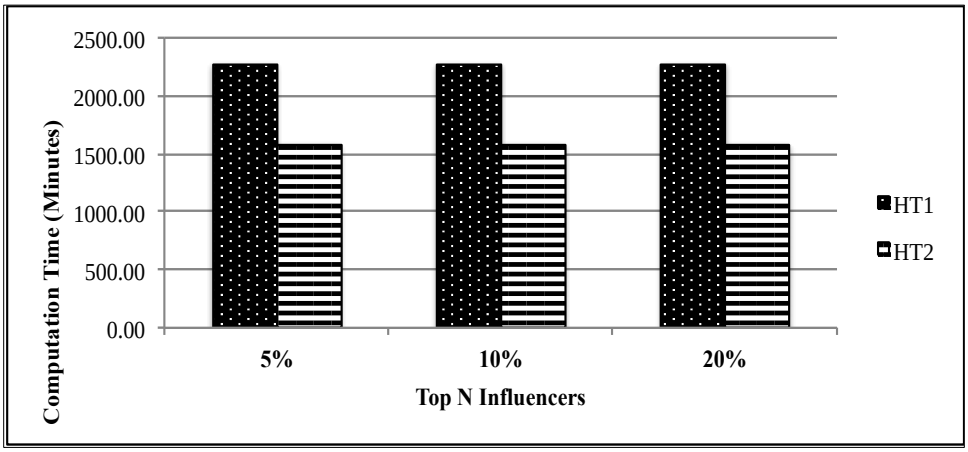
(c-1) HITS-based Algorithms, Window 3, *N*=2,510



Coverage Rate/ Computation Time

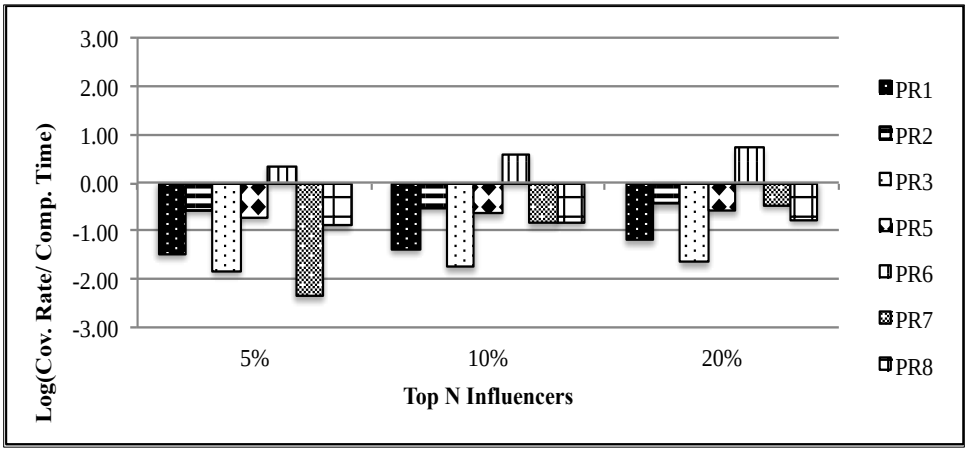


Coverage Rate

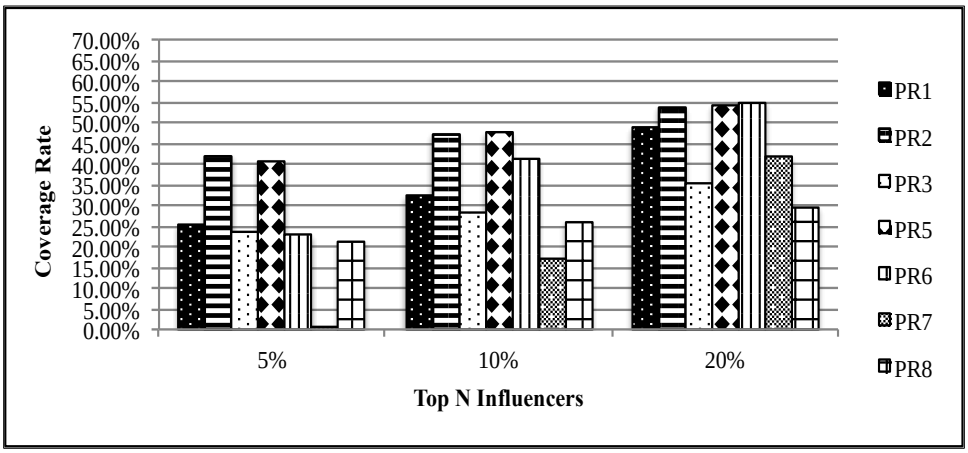


Computation Time

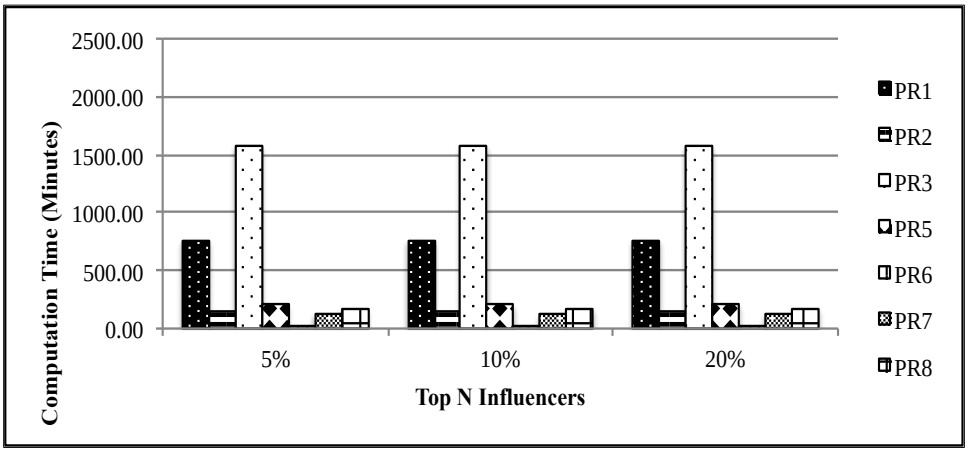
(c-2) PageRank-based Algorithms, Window 3, *N*=2,510



Coverage Rate/ Computation Time

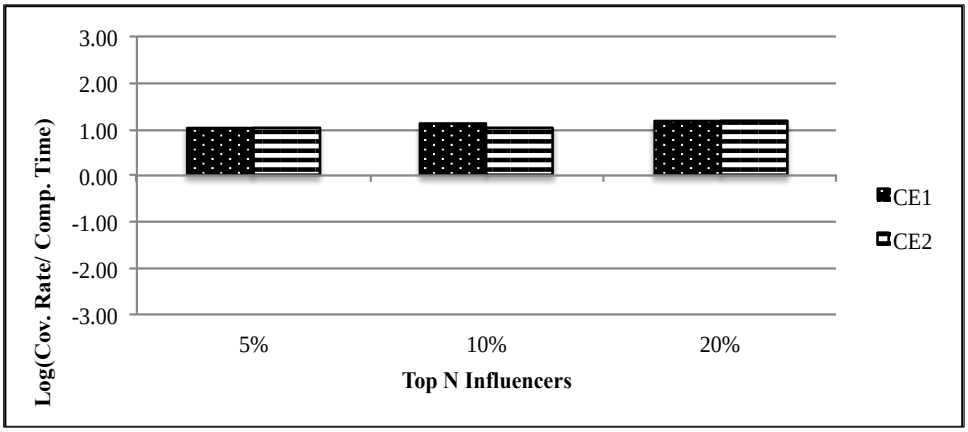


Coverage Rate

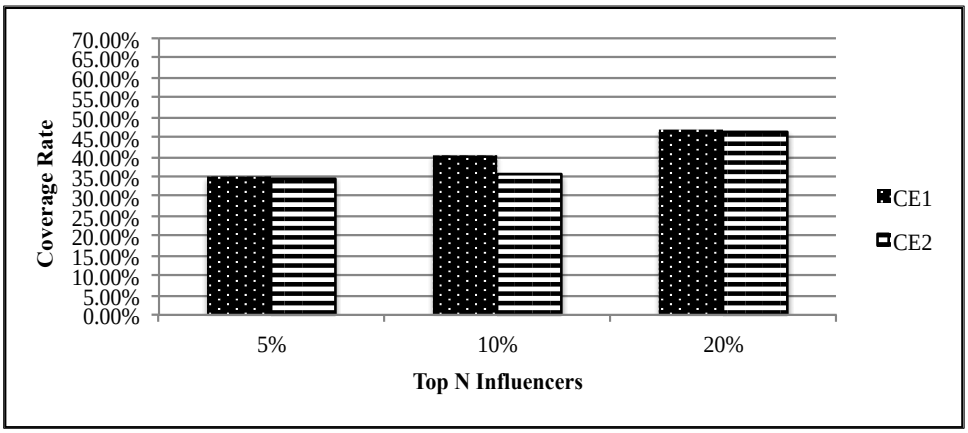


Computation Time

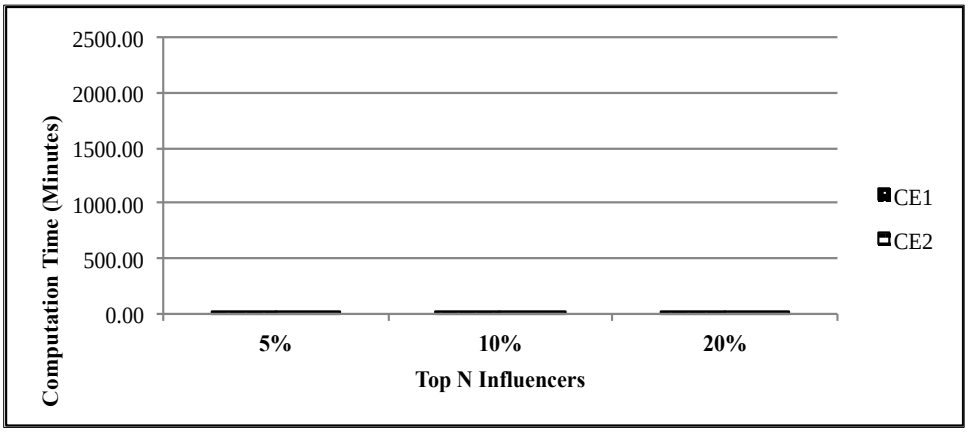
(c-3) Centrality-based Mechanisms, Window 3, *N*=2,510



Coverage Rate/ Computation Time

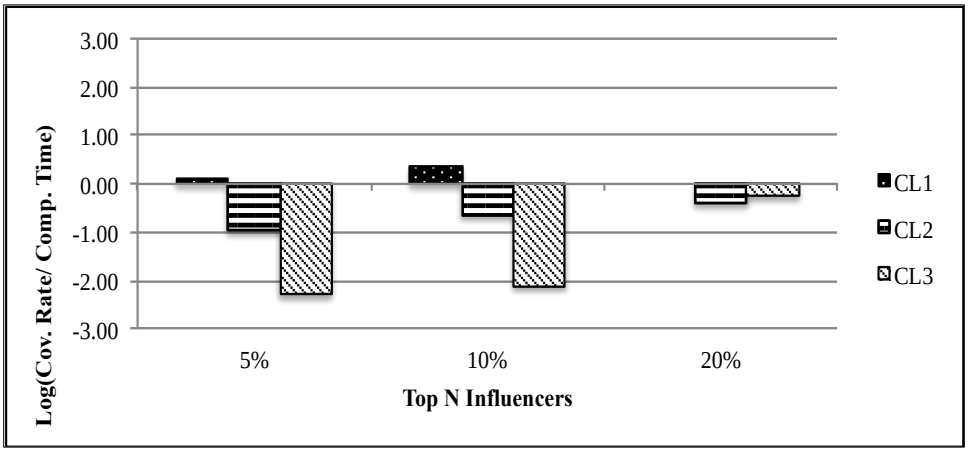


Coverage Rate

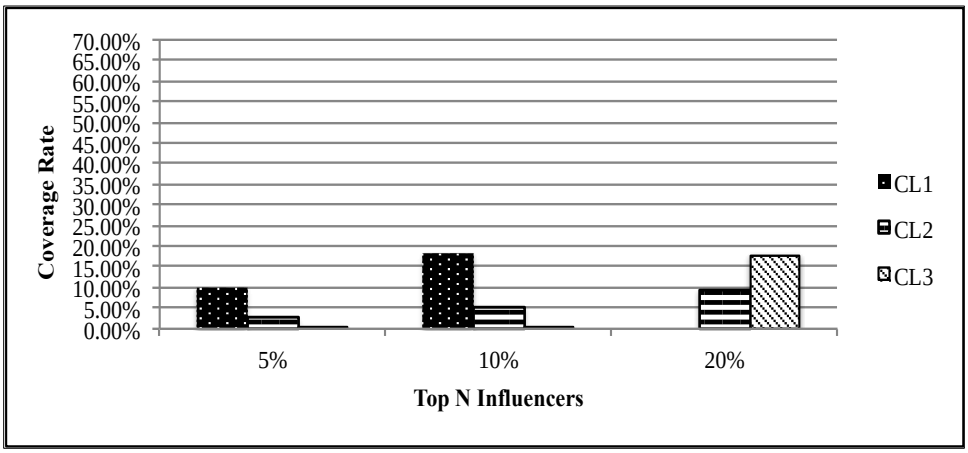


Computation Time

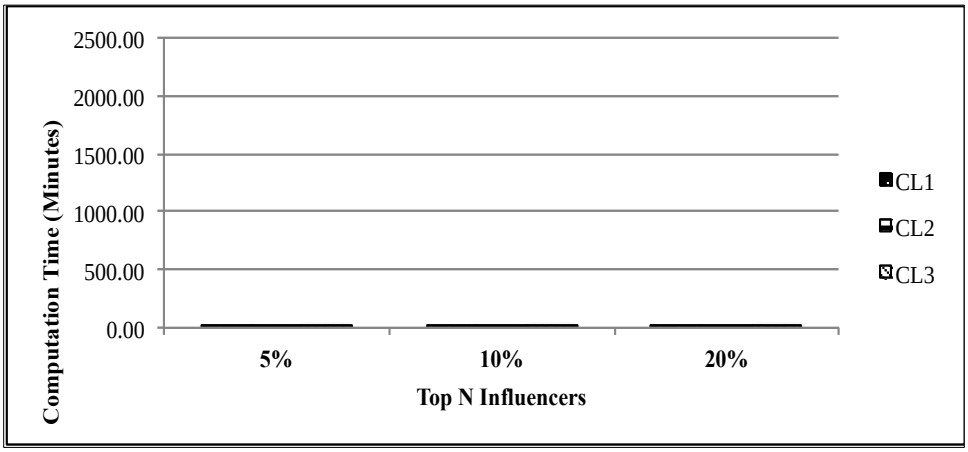
(c-4) Clustering-based Algorithms, Window 3, *N*=2,510



Coverage Rate/ Computation Time



Coverage Rate



Computation Time

Figure 3.8 shows that the influencers identified by the centrality-based methods (CE1 and CE2) have a better quality in terms of the bang-to-buck ratio between coverage rate and computation time. However, as the line charts in Figure 3.7 show, the centrality-based methods did not produce influencers with the best coverage rates among all the identification approaches. But because the computation time for centrality-based methods is relatively short, the coverage rate bang-to-buck ratio for influencers identified by the centrality-based methods is better than this bang-to-buck ratio for the other algorithms/methods.

Among the HITS-based algorithms, HT1 and HT2 produce two groups of influencers with similar quality in terms of the coverage rate / computation time ratio. In the PageRank-based algorithm group, even though PR2 and PR5 produce influencers with a high coverage rate, the long computation time for these two algorithms causes a low bang-to-buck ratio. PR6 identifies influencers with the best performance in terms of the coverage rate / computation time ratio, mainly because of the short computation time. In the clustering-based algorithm group, the bang-to-buck ratio for the influencers identified by CL1 is slightly better than those for CL2 and CL3.

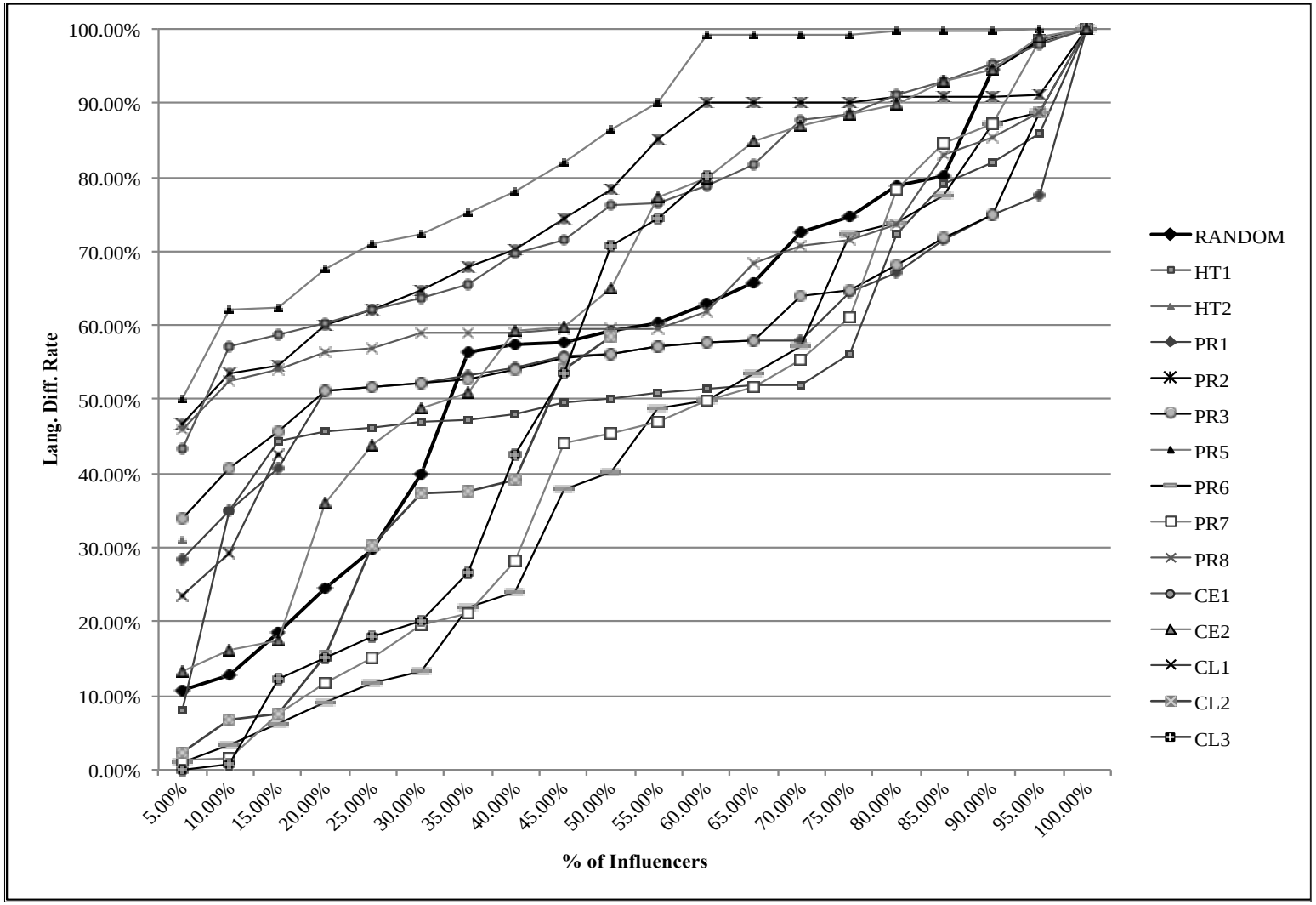
Figure 3.8 reveals that in the context of influencer identification, different metrics are useful in helping to select approaches based on the analysis requirements. If the dataset is relatively small and there is little time pressure, approaches that identify higher quality influencers but have long computation times may be preferable. When analyzing a huge network, the analyst should consider the computation time and make corresponding changes in the identification approach. When both time and the quality of influencers need to be taken into consideration simultaneously, the bang-to-buck ratio provides a good guideline for selecting an algorithm/method.

3.8.2.2 Language Diffusion Rate

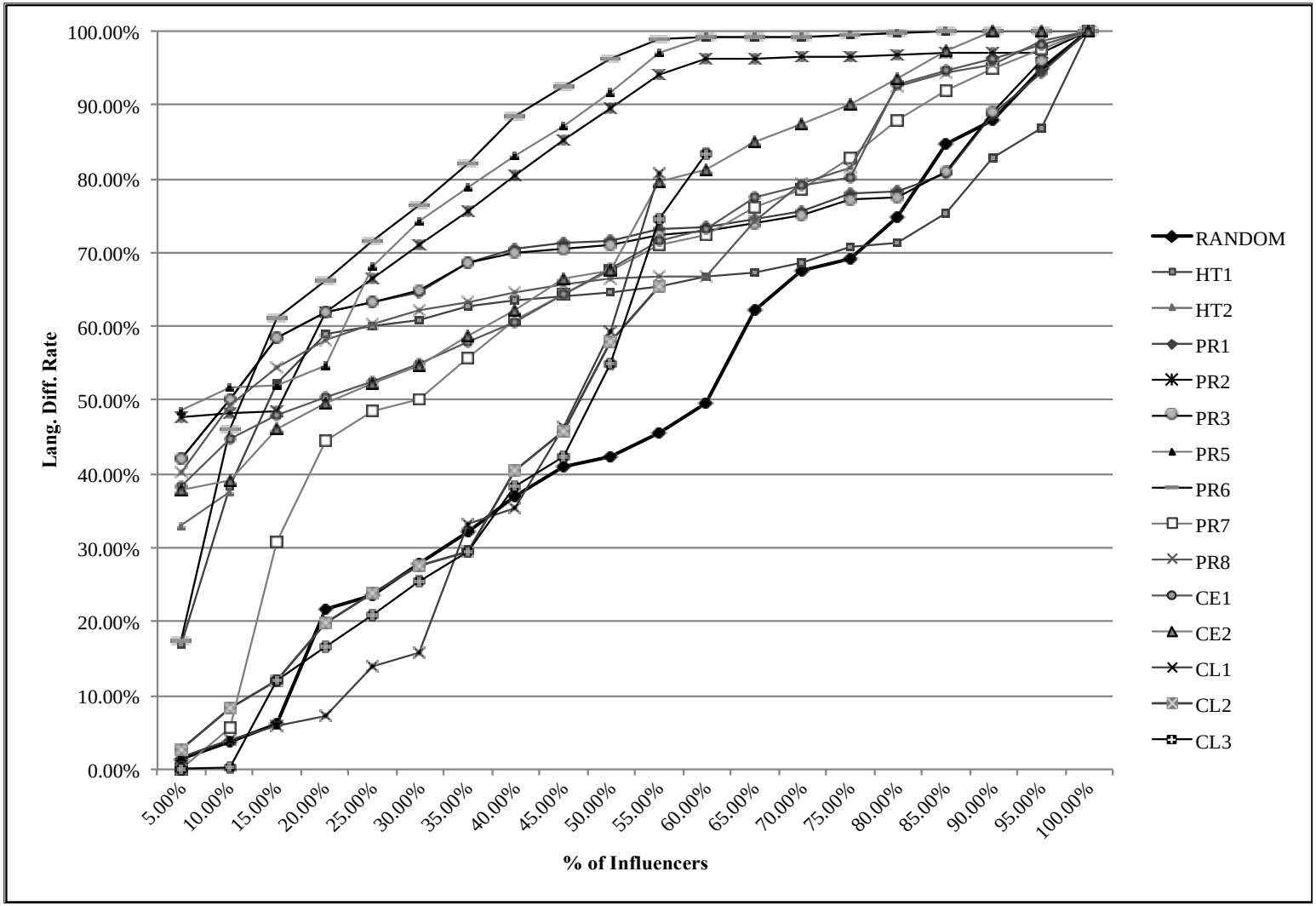
The foregoing results depend on the coverage rate metric as the sole quality. Here, consider another metric, the language diffusion rate, to show how the words used by the influencers transfer to those they cover. This metric involves how the information produced by influencers makes an impact during the interactions between influencers and other individuals. For example, if an influencer is talking about the election event in the state of Indiana and the covered others who reply to his/her posting mention the word *Indiana*, this means that the influencer and the covered others are discussing the same event, and this can be explained as an information transfer from the influencer to the covered others.

Figure 3.9 Language Diffusion Rate for Different Algorithms: Twitter Election Dataset

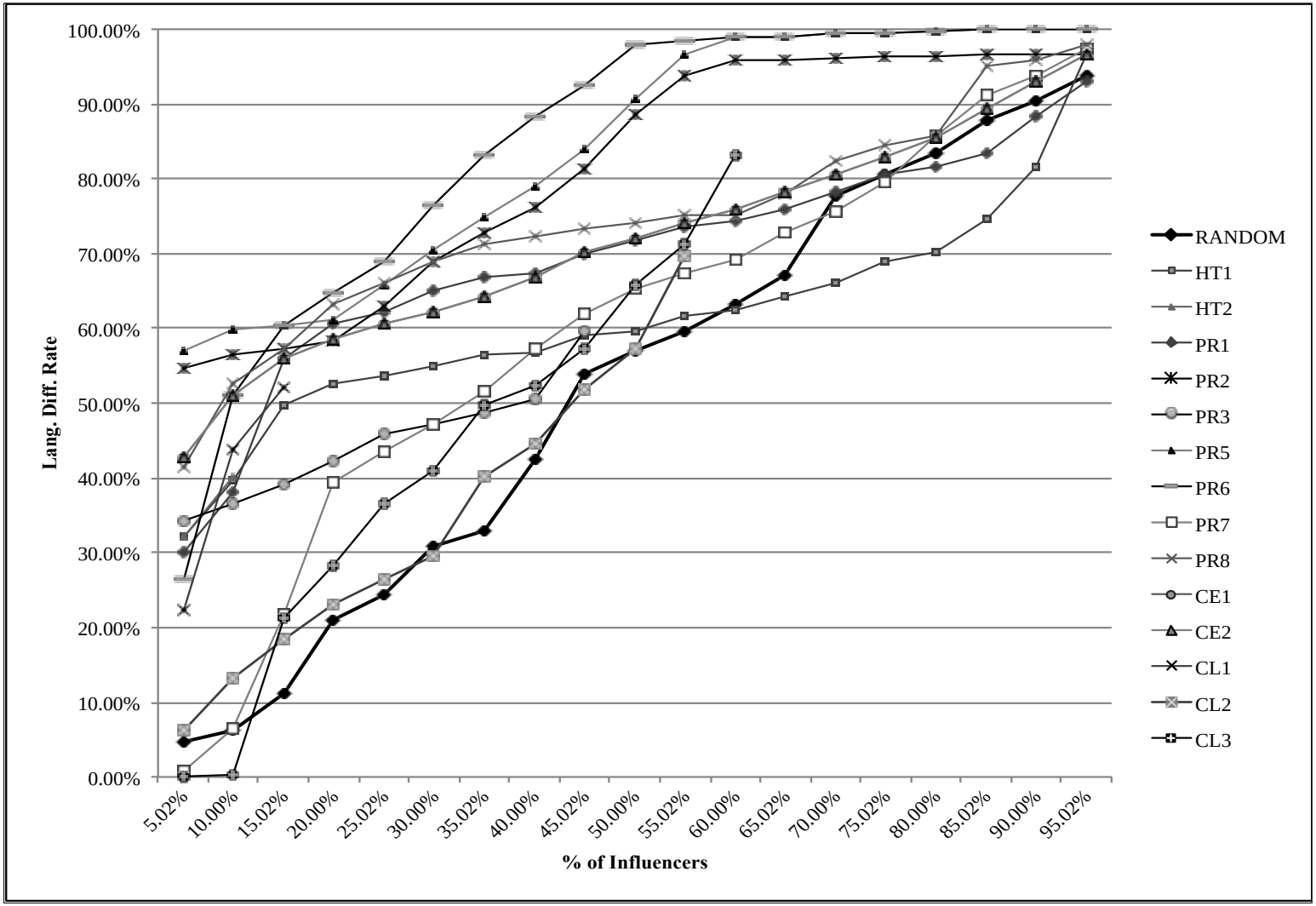
(a) Window 1, *N*=720



(b) Window 2, *N*=1,700



(c) Window 3, *N*=2,510

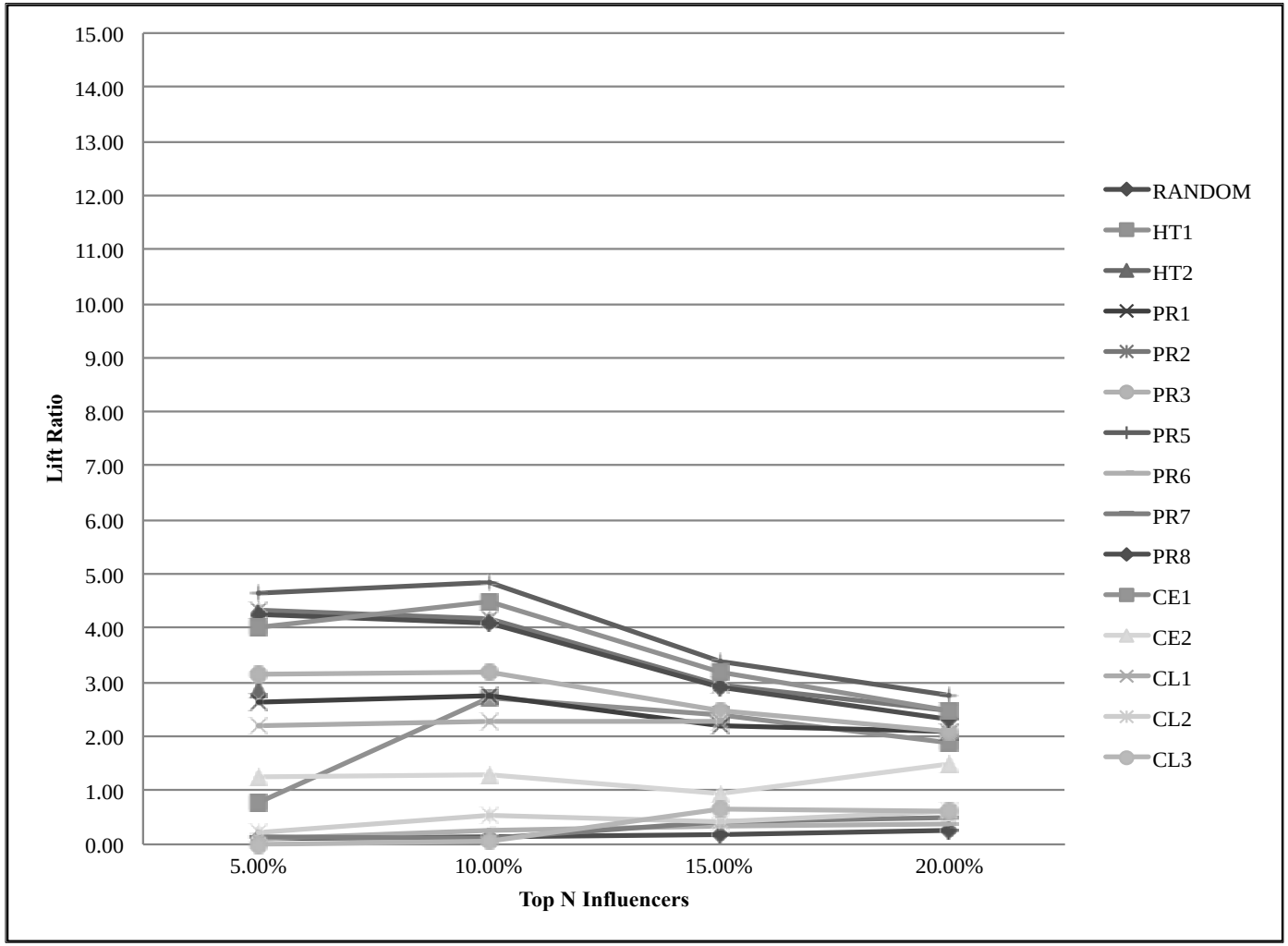


Overall, as shown in Figure 3.9, the influencers identified by PR2, PR5, and PR6 have the best language diffusion rates across all the approaches. The influencers identified by CE1 and CE2 also show a good language diffusion rate. These results are similar to those for the coverage metrics, which means that methods can identify those influencers with high coverage rates and also strong language diffusion rates.

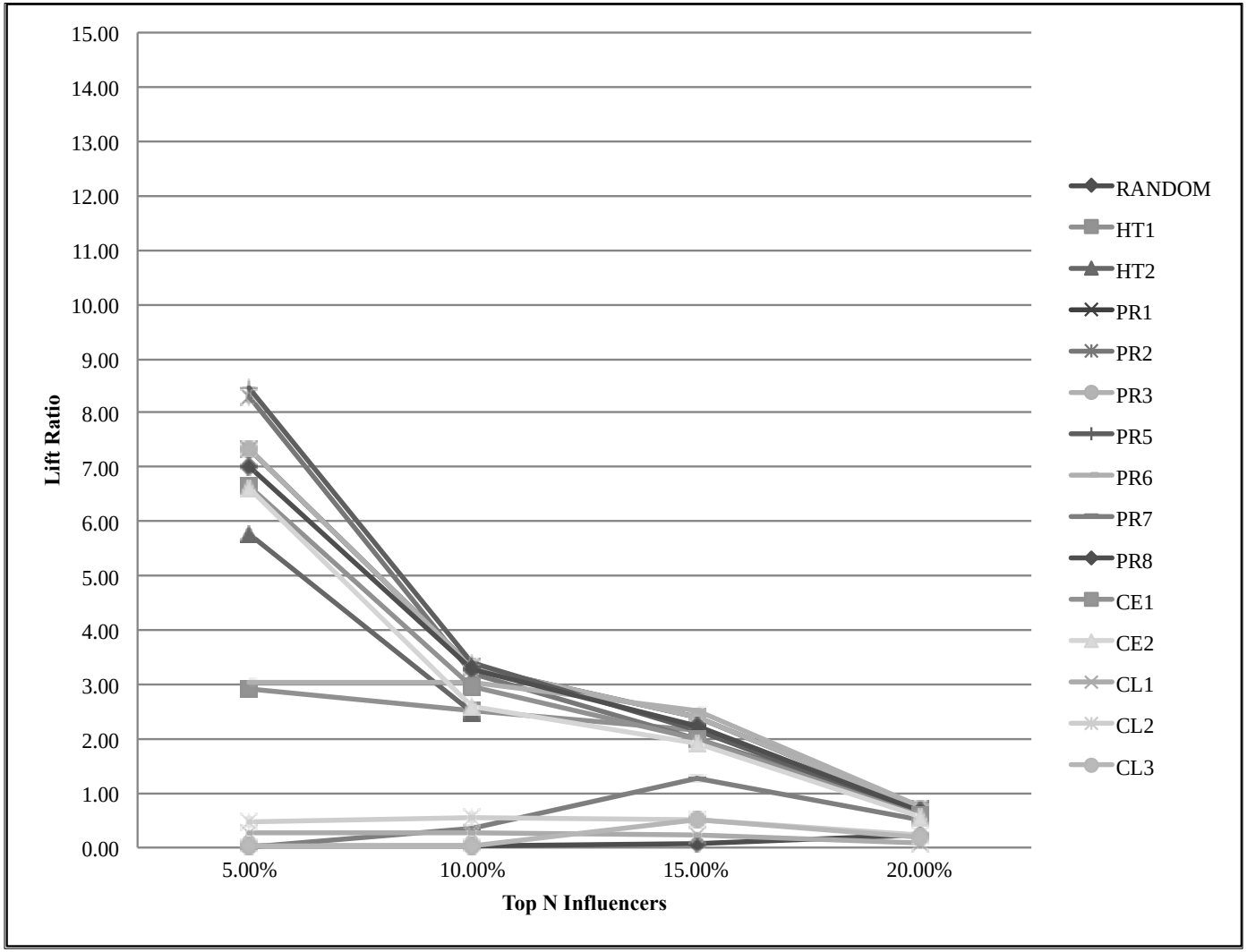
In contrast, the case of PR8 presents a difference in the qualities of the language diffusion rate for the influencers it has identified, compared to the quality of their coverage rate. This means that the influencers identified by PR8 would reach many people, but these people would not diffuse the influencers' words. The Figure 3.10 lift ratio charts provide a closer look at varied language diffusion rates within the top 20% of identified influencers.

Figure 3.10 Language Diffusion Rate Lift Ratio Charts for Different Algorithms: Twitter Election Dataset

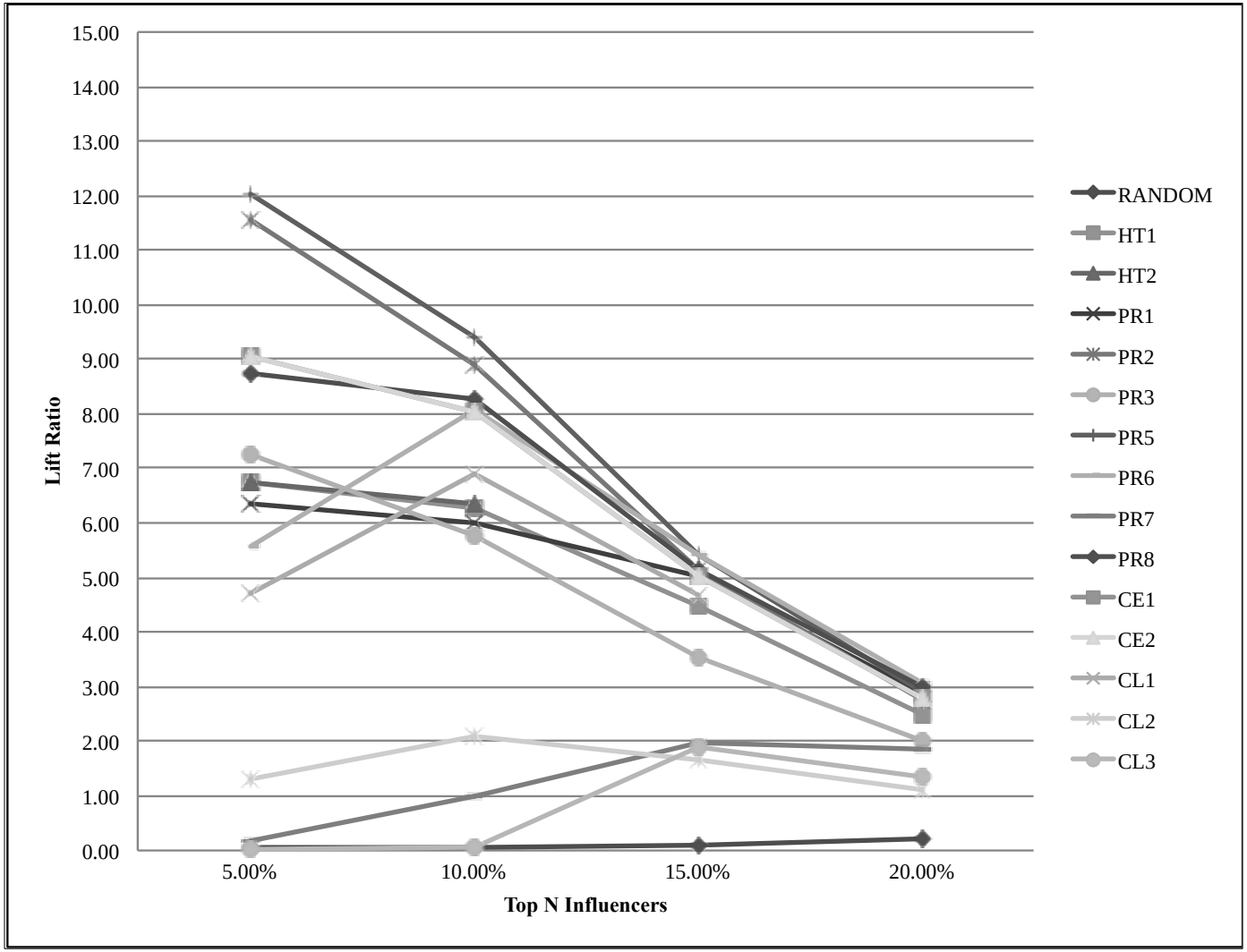
(a) Window 1, *N*=720



(b) Window 2, *N*=1,700



(c) Window 3, *N*=2,510

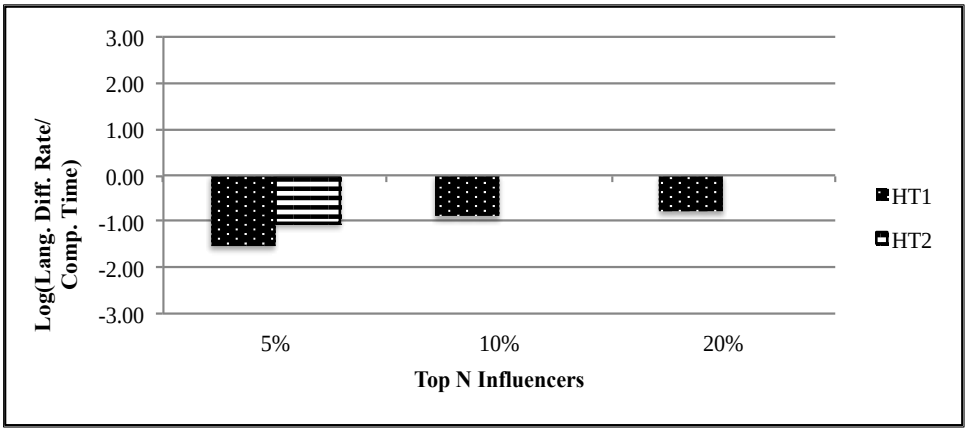


Looking at the top 20% of identified influencers, PR2, PR5, and PR6 found influencers with a superior quality language diffusion rate than the others. These results are similar to the results for the coverage rate metrics. In contrast, the influencers identified by CE1 and CE2 have medium to low rates of language diffusion. The lift ratio charts also show the similar results to the ones in coverage rate. All the lines are getting closer when the percentage of identified influencers increases. This means when identifying small group of influencers, language diffusion also shows very differently in different approach.

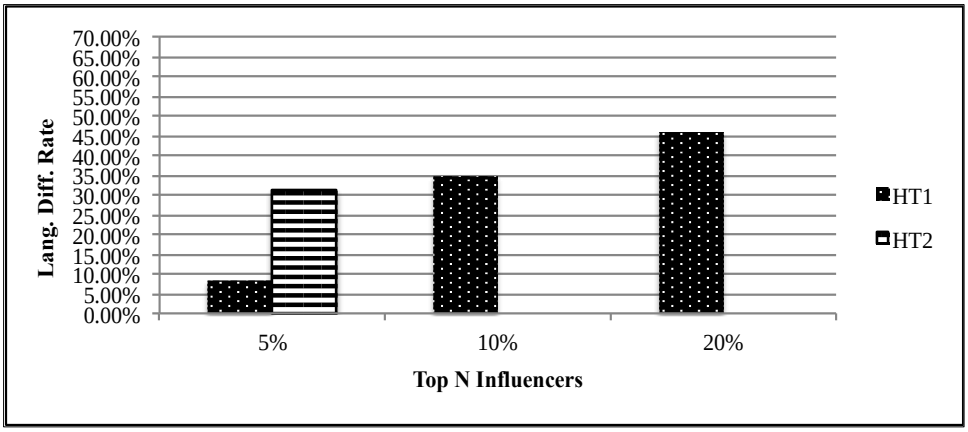
The charts in Figure 3.11 shows the bang-to-buck ratios between the language diffusion rates and computation times. The line charts are mainly for presenting the trend of each approach in identifying small to larger groups of influencers. The bang-to-buck charts show the quality of the different approaches when considering both computation time and the language diffusion rate. These charts can be compared to those for the coverage rates to see the differences between the results for these two metrics.

Figure 3.11 Language Diffusion Rate of Top *N*% of Influencers for Different Algorithms in Different Categories: Twitter Election Dataset

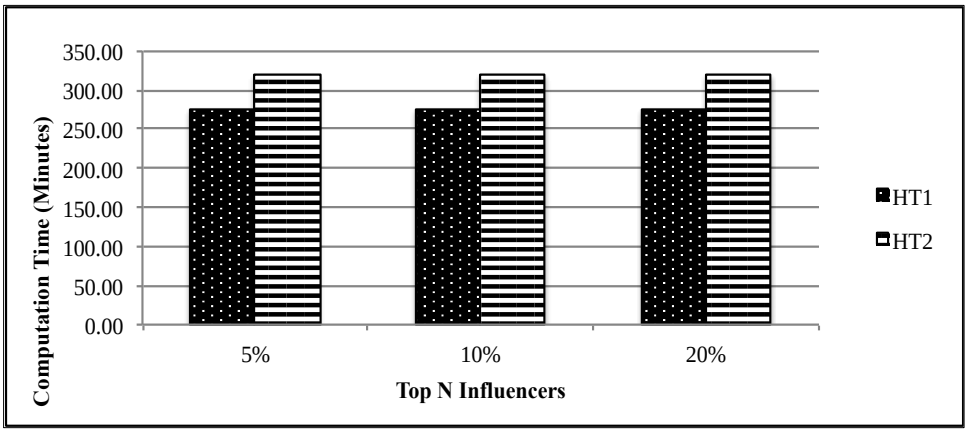
(a-1) HITS-based Algorithms, Window 1, *N*=720



Coverage Rate/ Computation Time

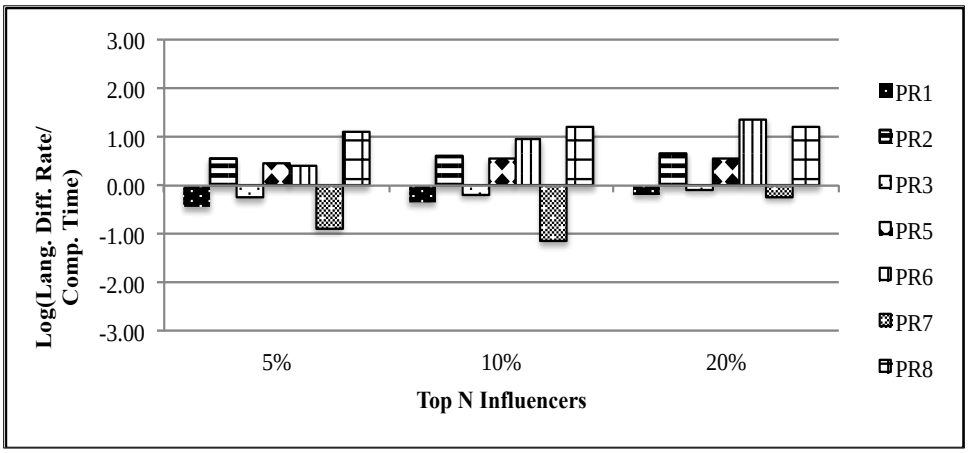


Language Diffusion Rate

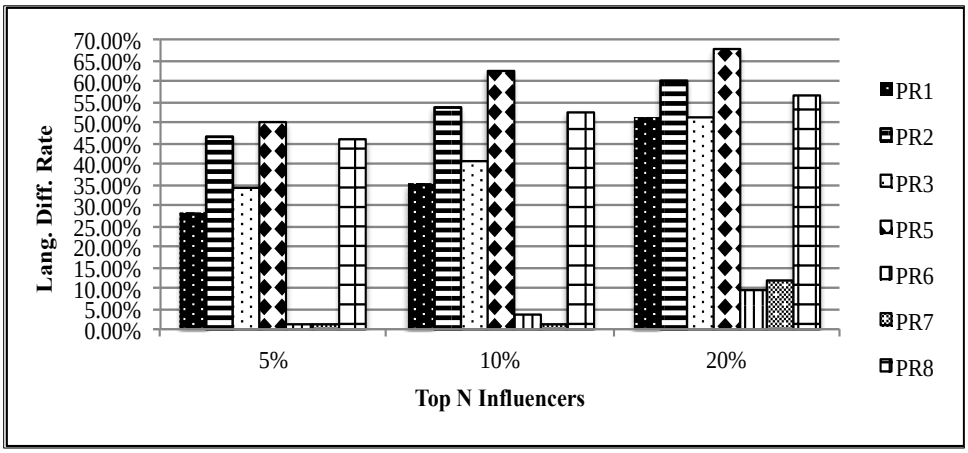


Computation Time

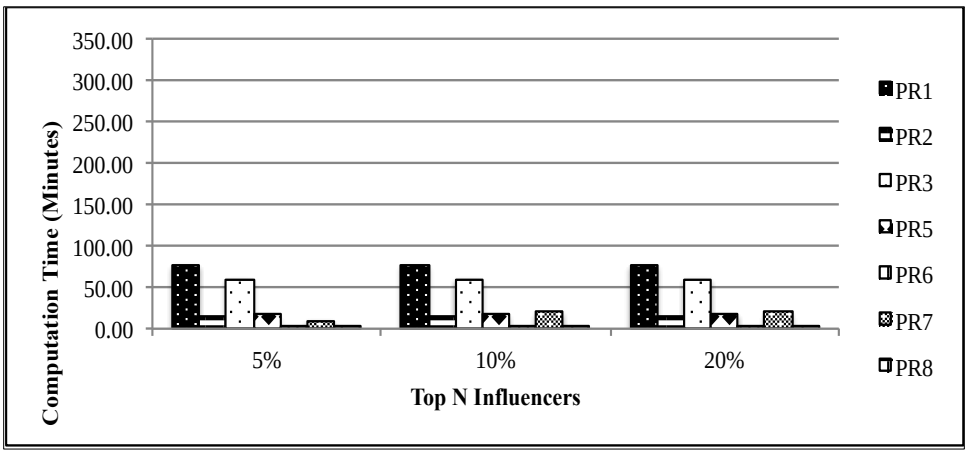
(a-2) PageRank-based Algorithms, Window 1, *N*=720



Coverage Rate/ Computation Time for

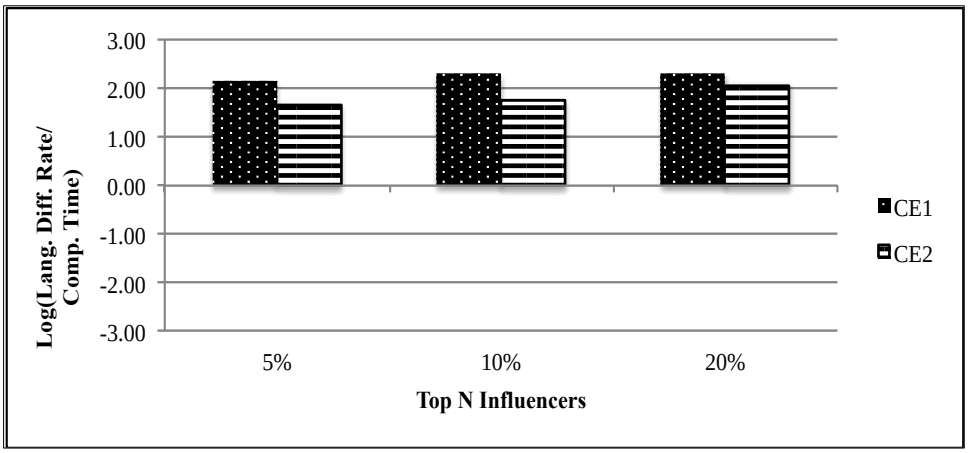


Language Diffusion Rate

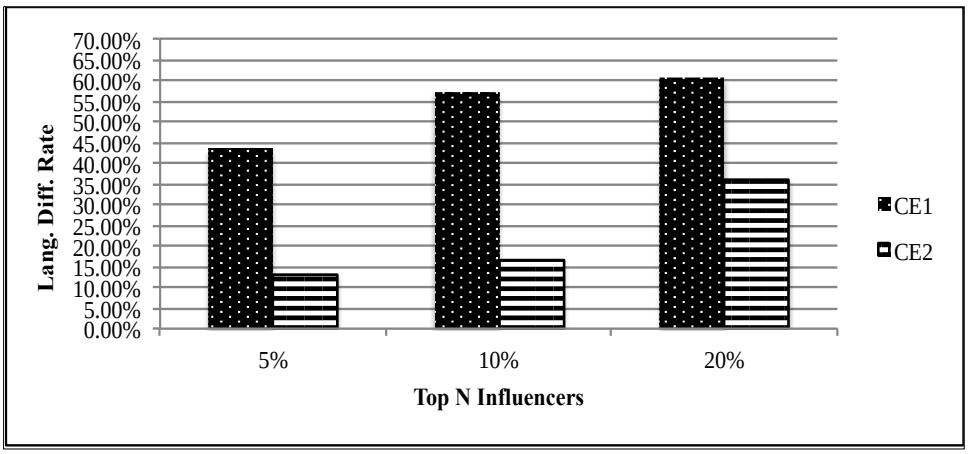


Computation Time

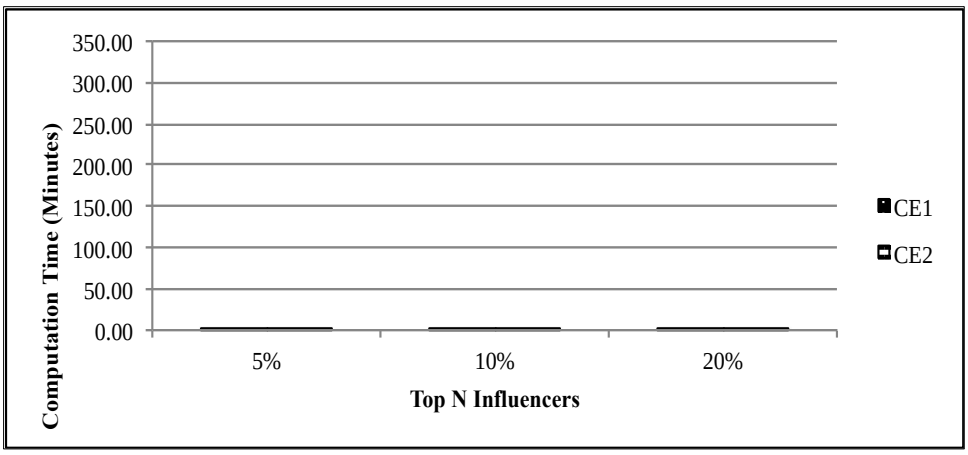
(a-3) Centrality-based Mechanisms, Window 1, *N*=720



Coverage Rate/ Computation Time

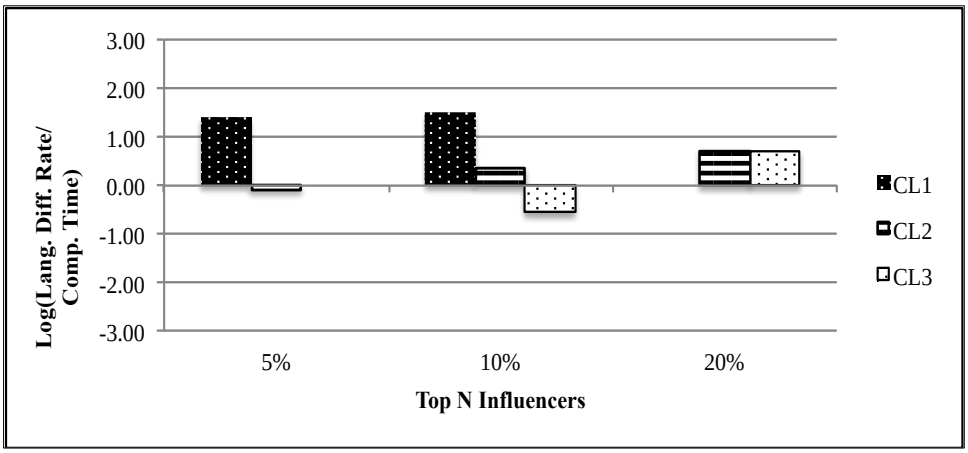


Language Diffusion Rate

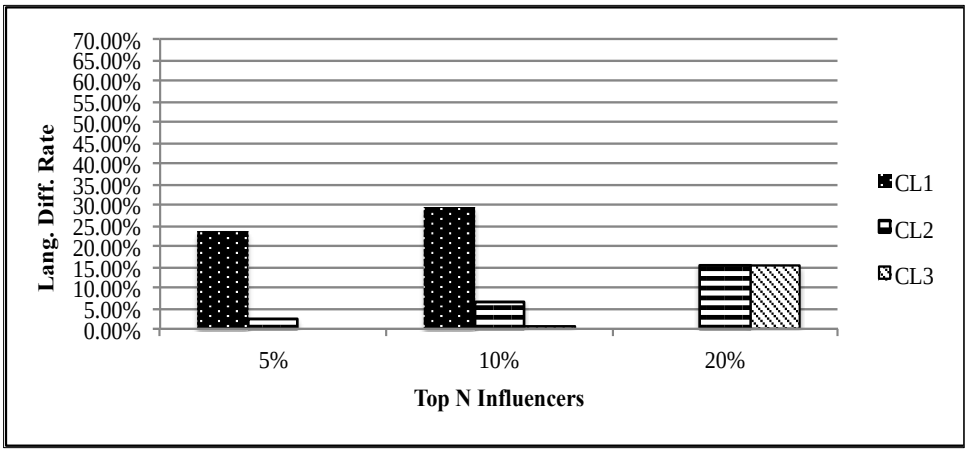


Computation Time

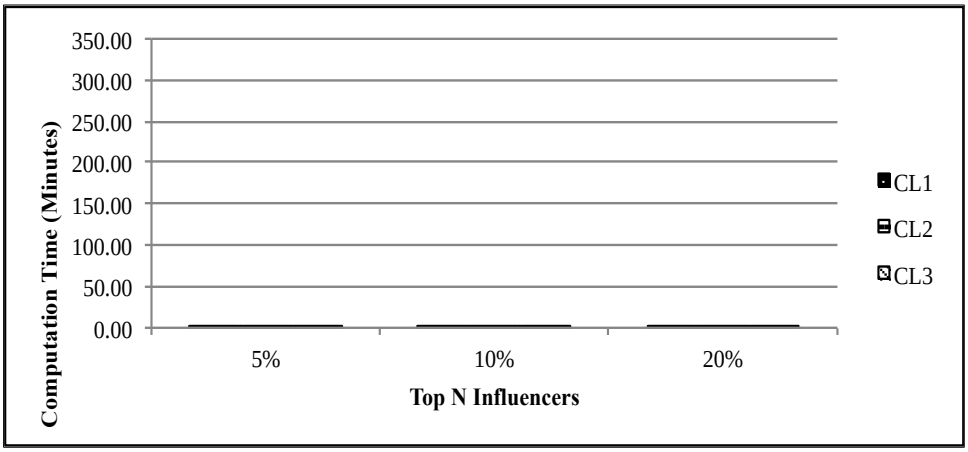
(a-4) Clustering-based Algorithms, Window 1, *N*=720



Coverage Rate/ Computation Time

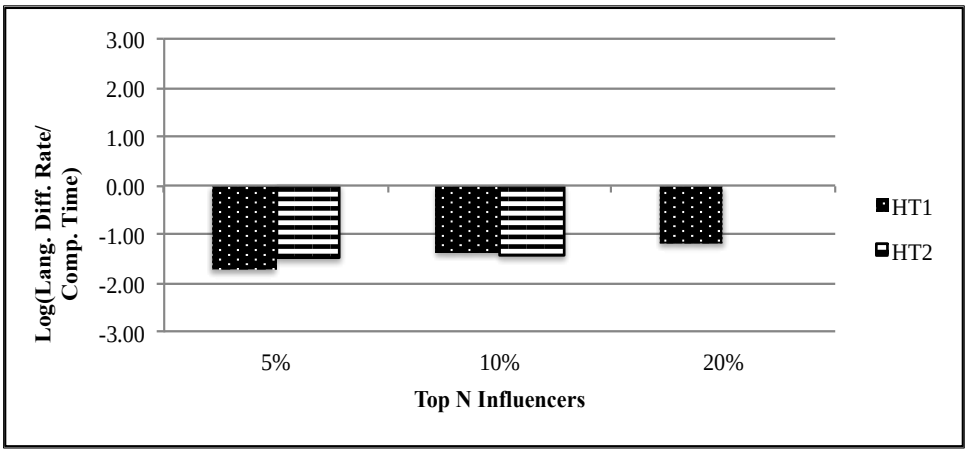


Language Diffusion Rate

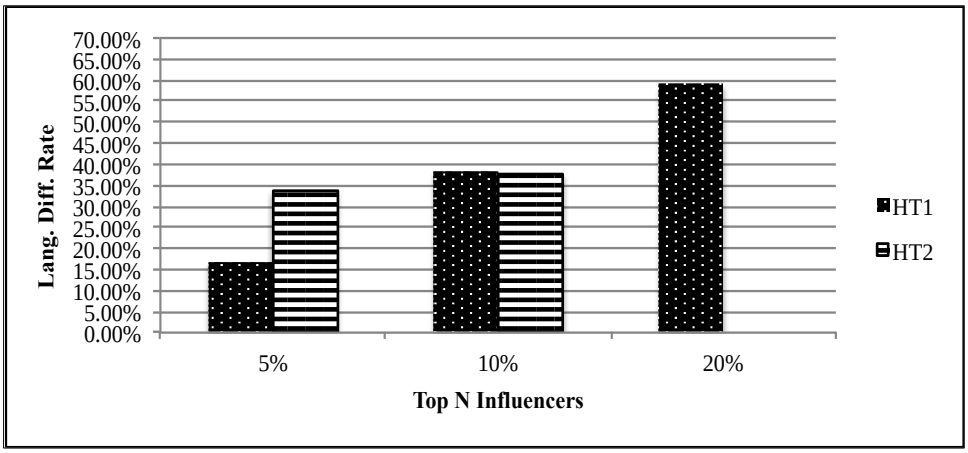


Computation Time

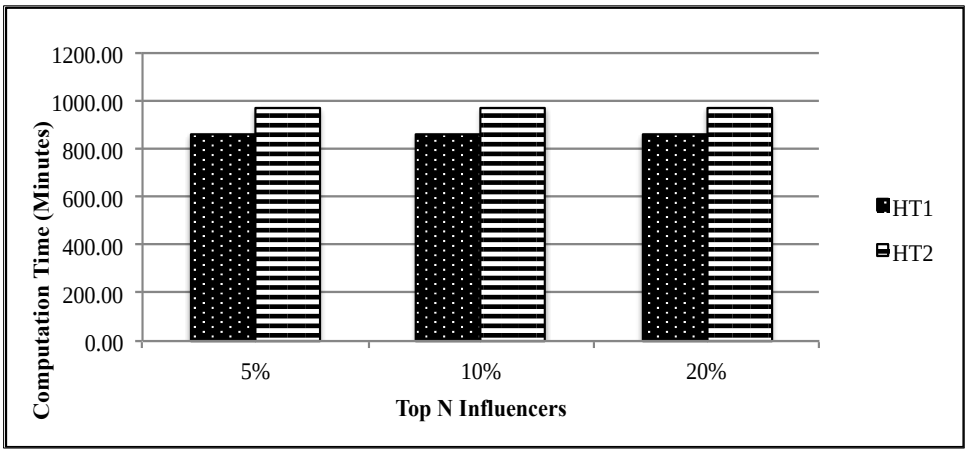
(b-1) HITS-based Algorithms, Window 2, *N*=1,700



Coverage Rate/ Computation Time

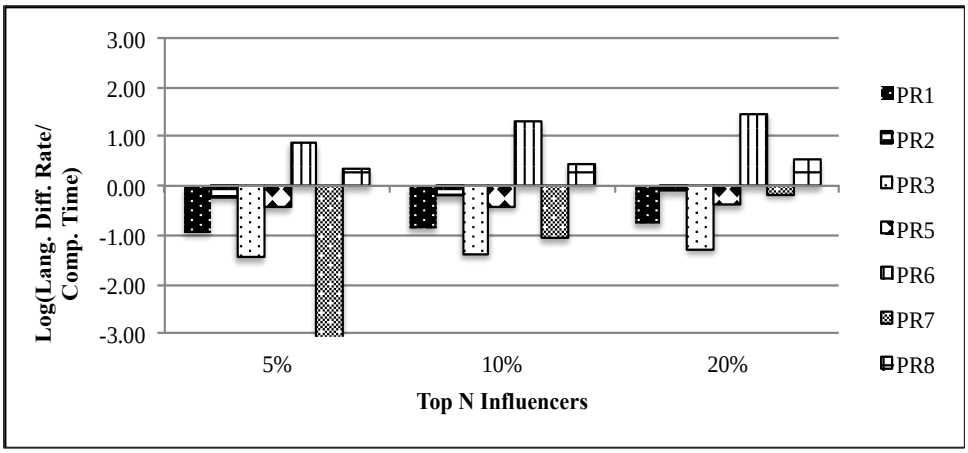


Language Diffusion Rate

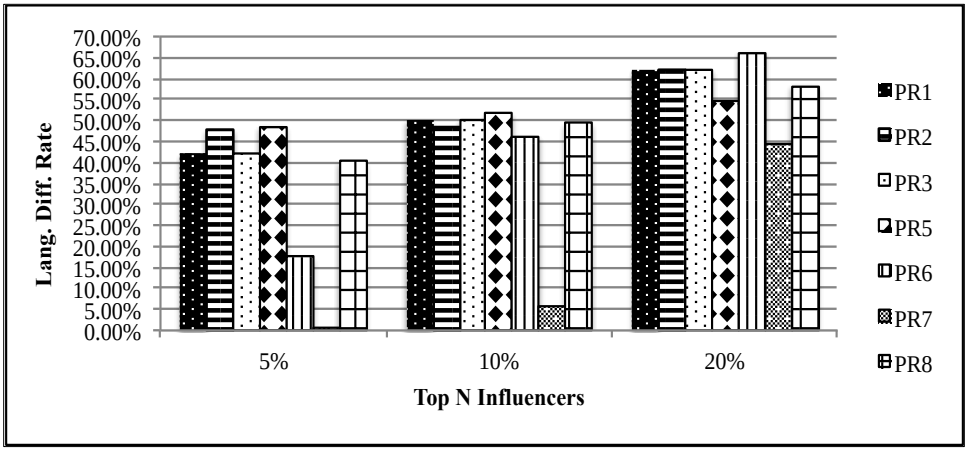


Computation Time

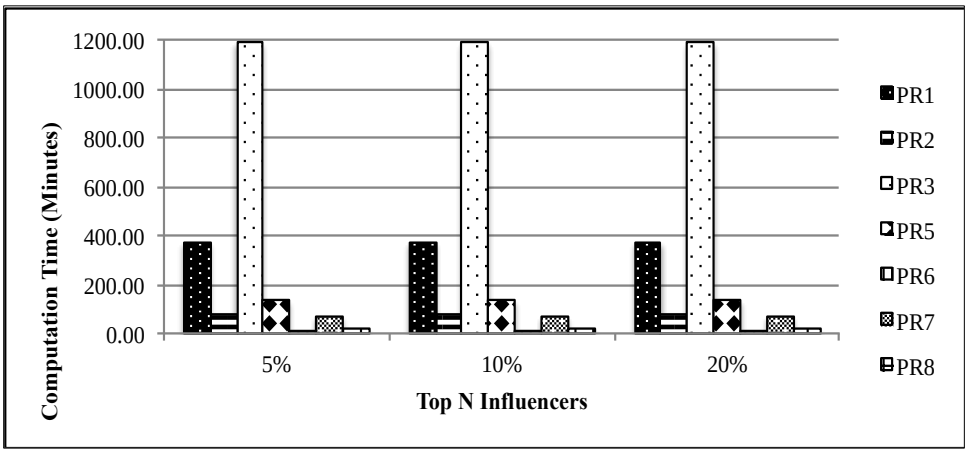
(b-2) PageRank-based Algorithms, Window 2, *N*=1,700



Coverage Rate/ Computation Time

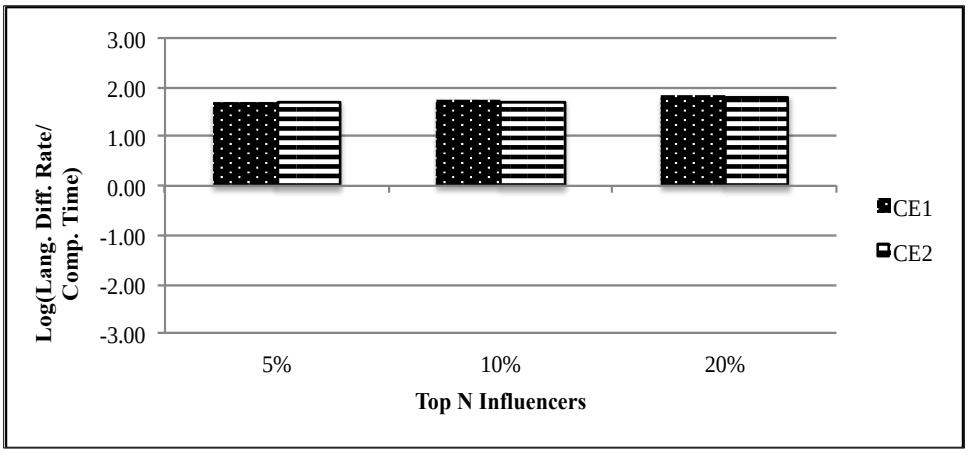


Language Diffusion Rate

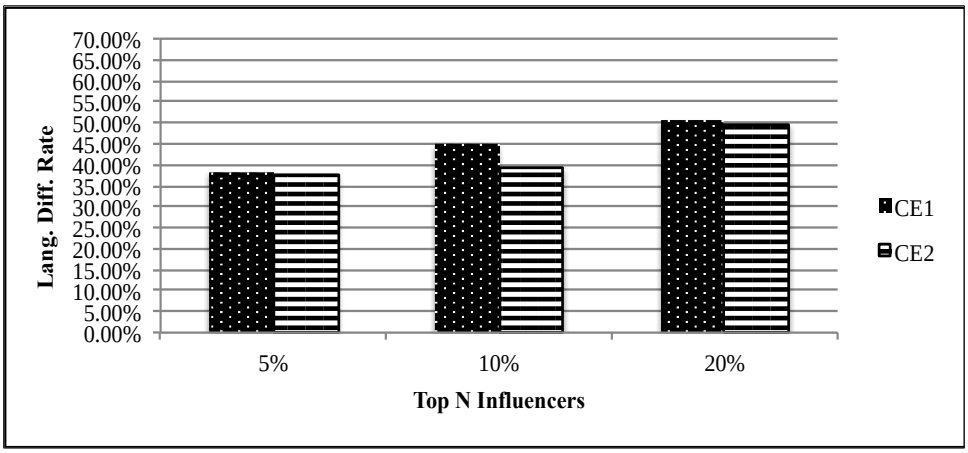


Computation Time

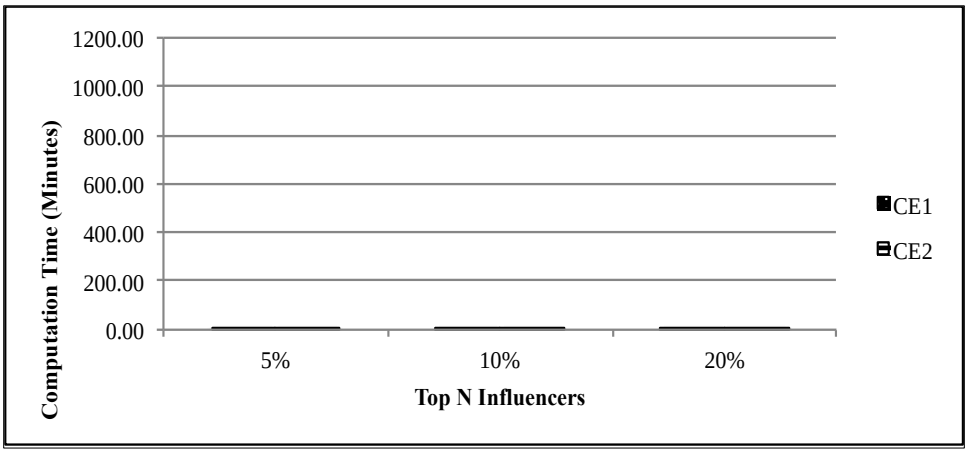
(b-3) Centrality-based Mechanisms, Window 2, *N*=1,700



Coverage Rate/ Computation Time

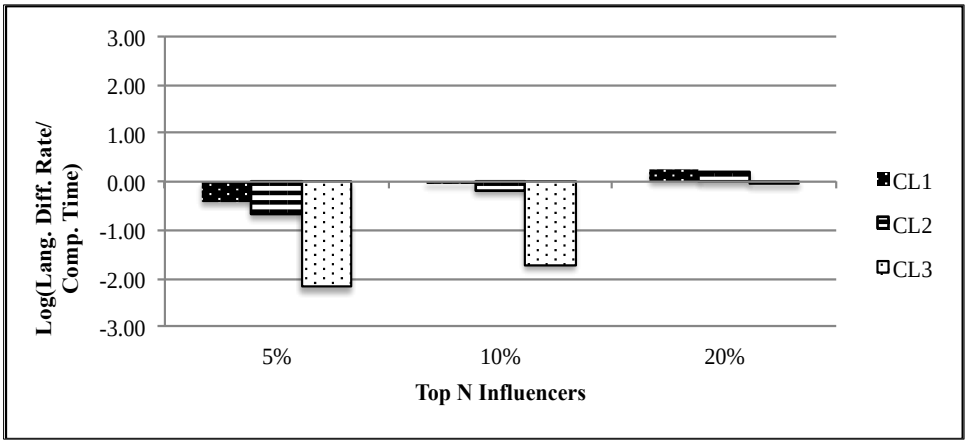


Language Diffusion Rate

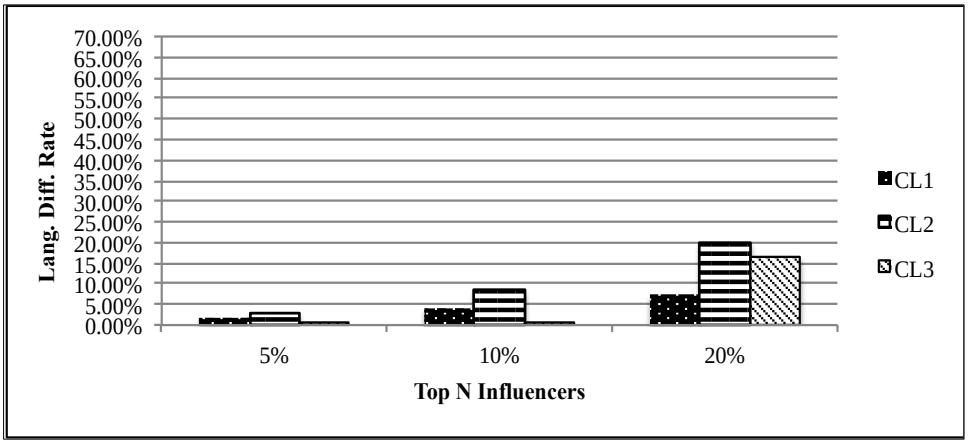


Computation Time

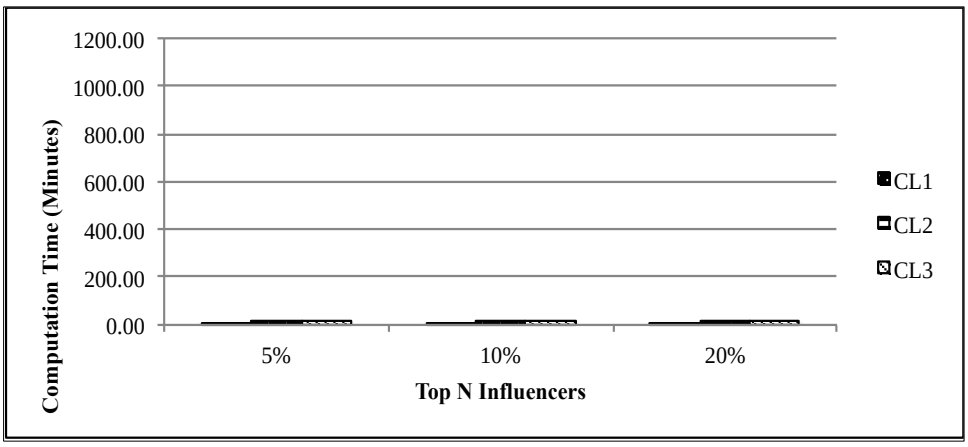
(b-4) Clustering-based Algorithms, Window 2, *N*=1,700



Coverage Rate/ Computation Time

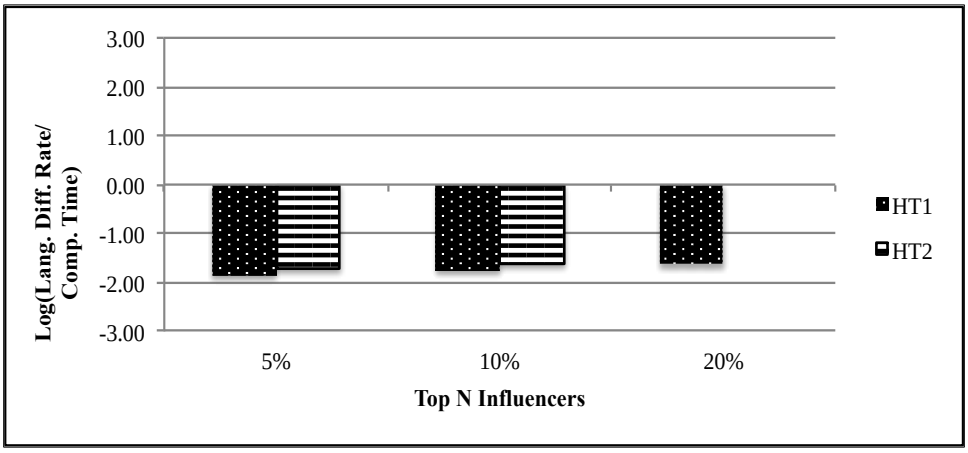


Language Diffusion Rate

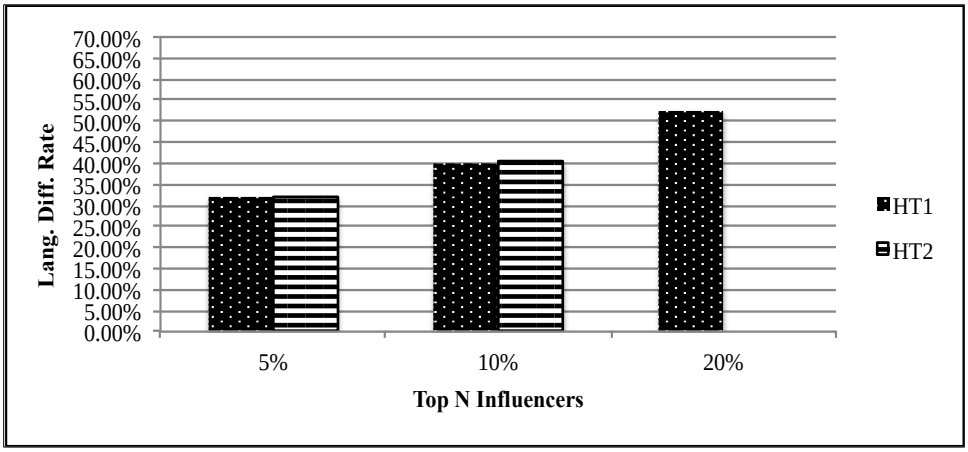


Computation Time

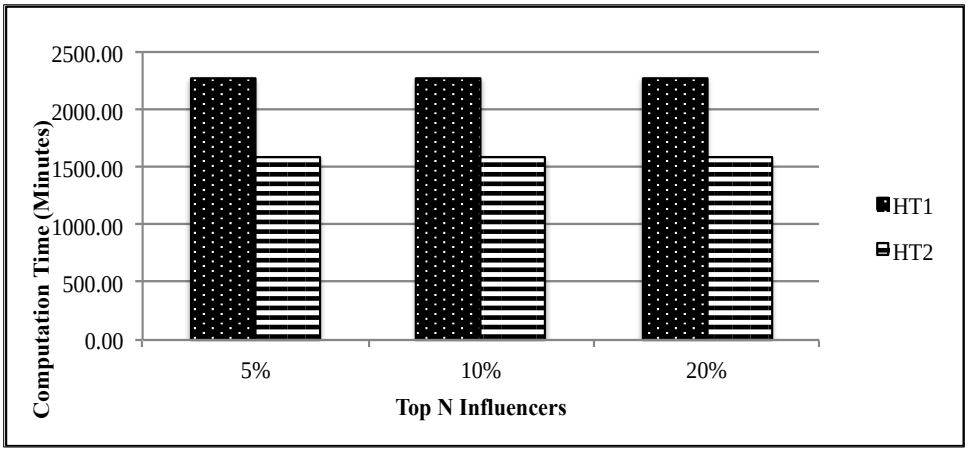
(c-1) HITS-based Algorithms, Window 3, *N*=2,510



Coverage Rate/ Computation Time

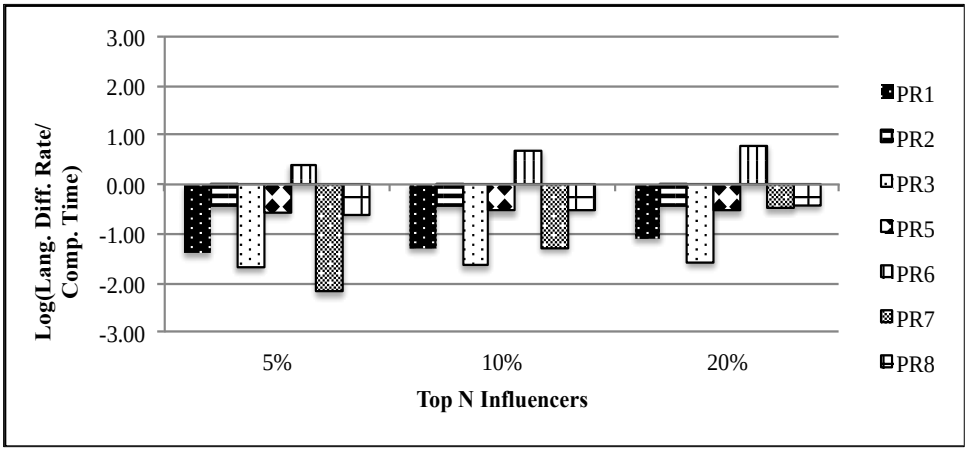


Language Diffusion Rate

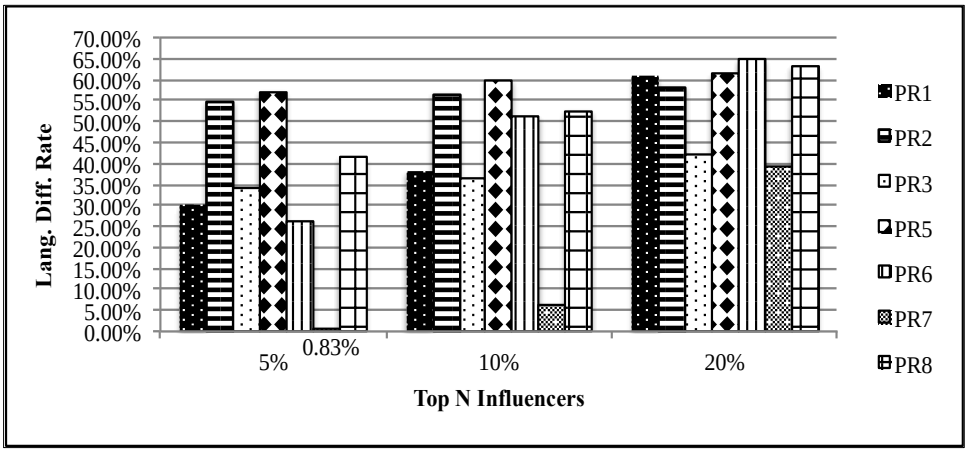


Computation Time

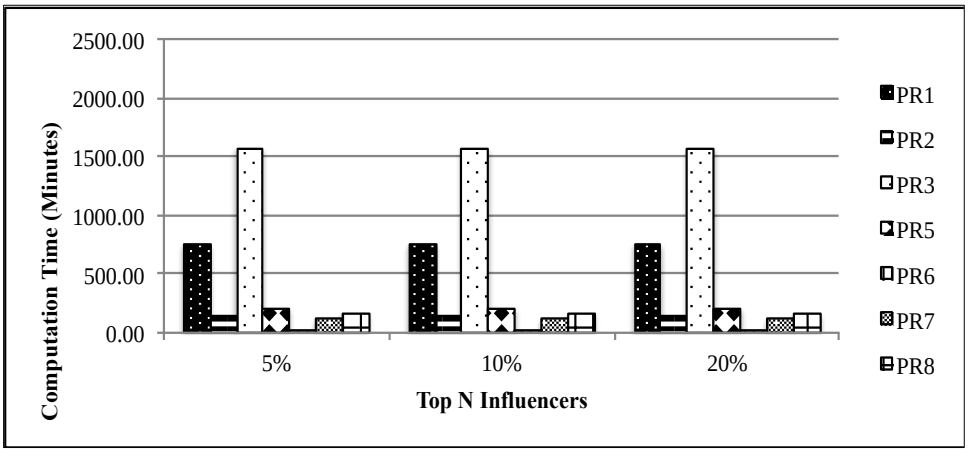
(c-2) PageRank-based Algorithms, Window 3, *N*=2,510



Coverage Rate/ Computation Time

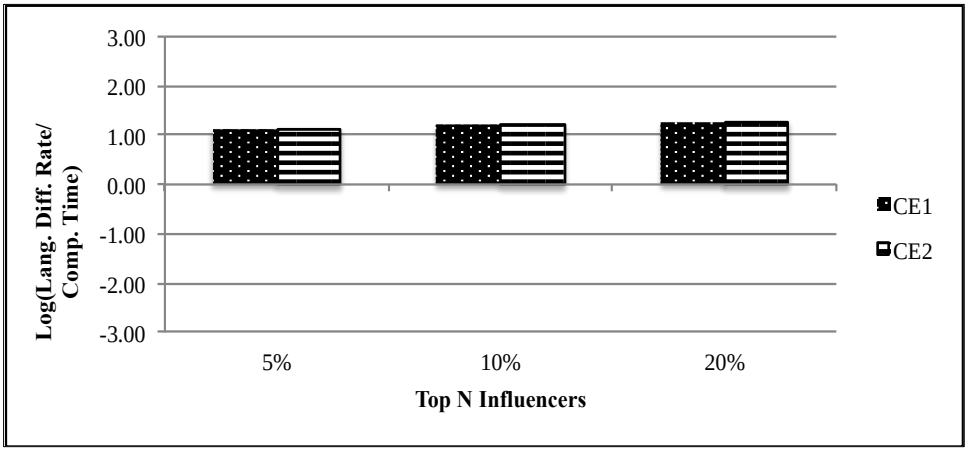


Language Diffusion Rate

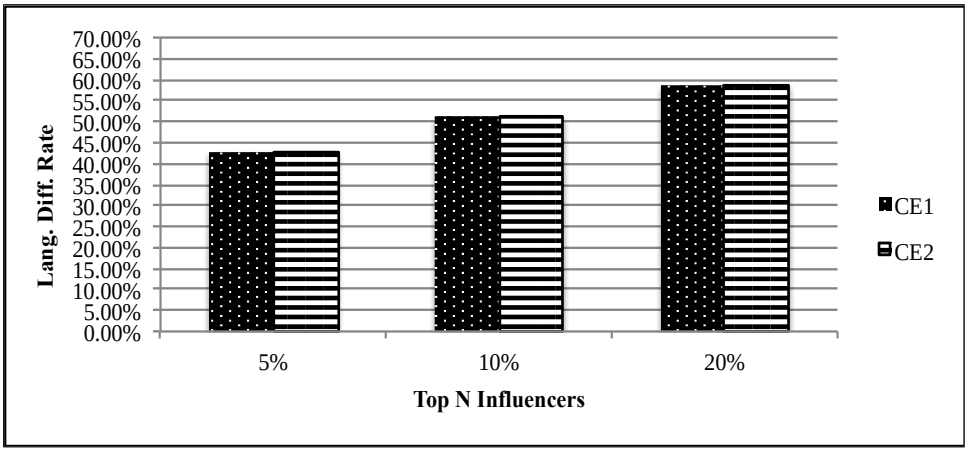


Computation Time

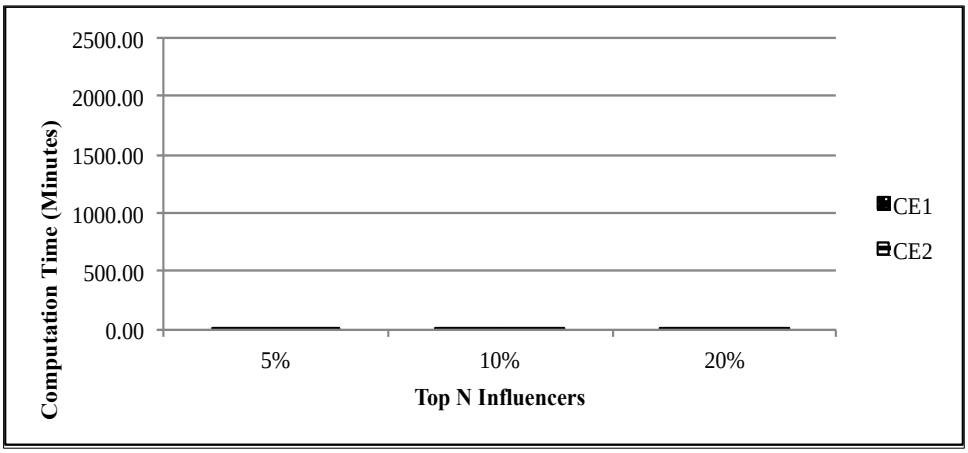
(c-3) Centrality-based Mechanisms, Window 3, *N*=2,510



Coverage Rate/ Computation Time

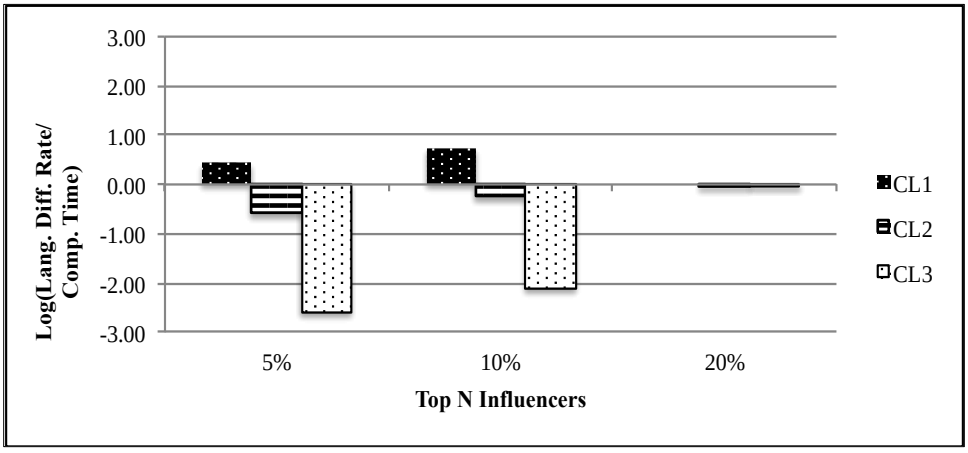


Language Diffusion Rate

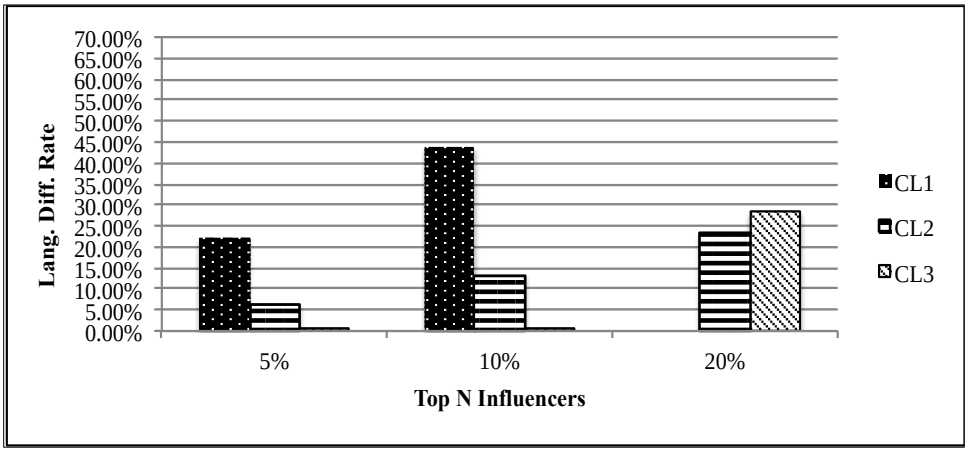


Computation Time

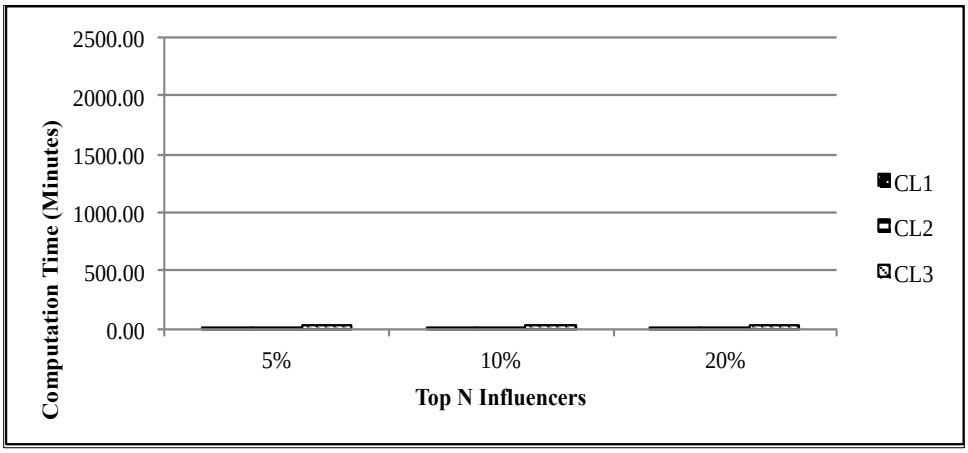
(c-4) Clustering-based Algorithms, Window 3, *N*=2,510



Coverage Rate/ Computation Time



Language Diffusion Rate



Computation Time

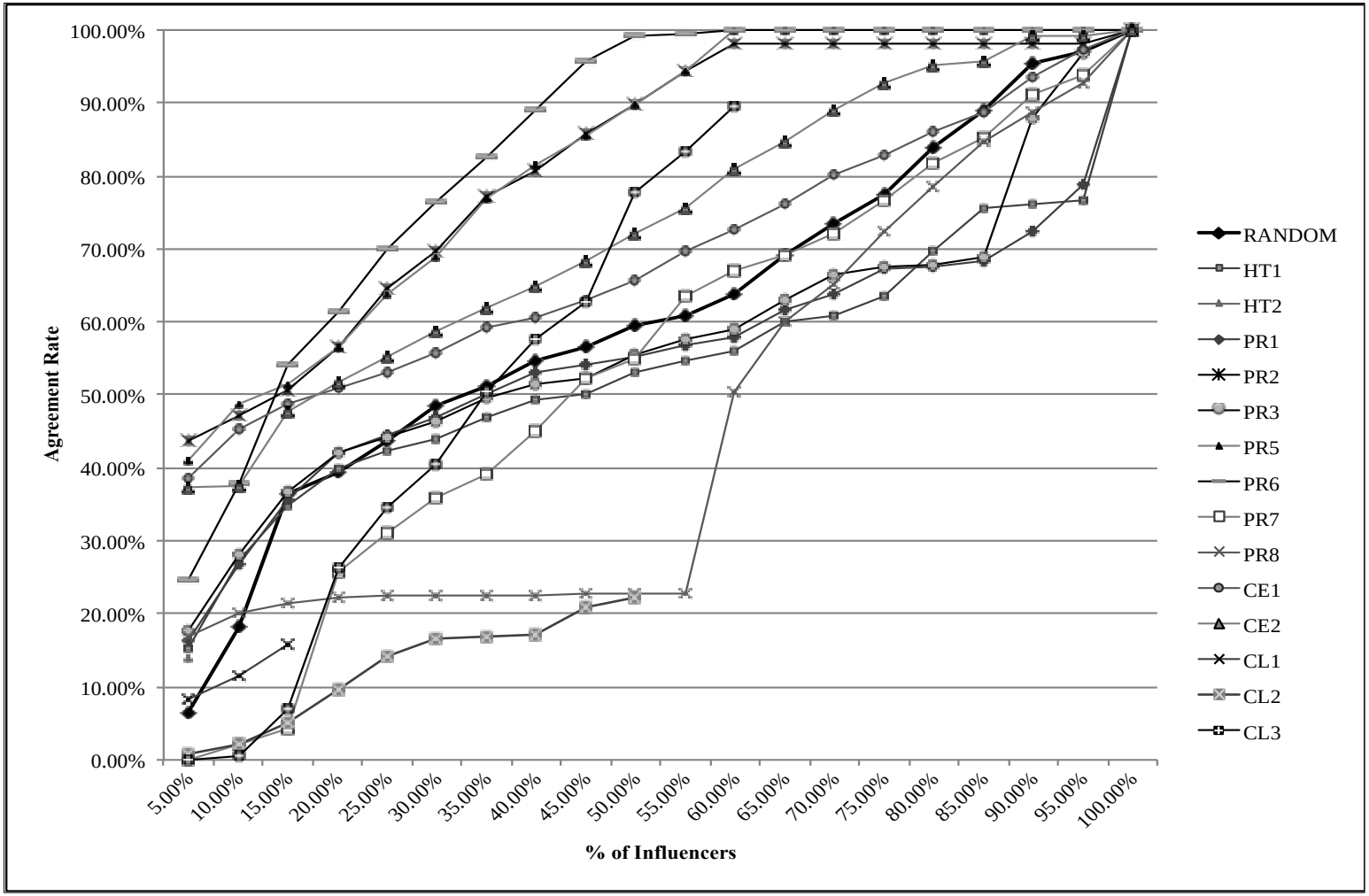
These results shown in these charts are similar to the results for the coverage rate. Among the PageRank-based algorithms, the influencers identified by PR6 also have the best language diffusion rate/computation time ratio. The influencers identified via CE1 and CE2 still yield the top quality in the ratio of language diffusion rate to computation time compared to other approaches. Because CE1 and CE2 take relatively little time for their computations, the identified influencers have a medium- to low-quality language diffusion rate, but are still the better selection in terms of the bang-to-buck ratio. In the group of HITS-based algorithms, the influencers identified by HT1 and HT2 show insignificant differences in the bang-to-buck ratio. As with the results shown in earlier charts, the influencers from CL1 produce the best quality among the three clustering algorithms.

3.8.2.3 Agreement Rate

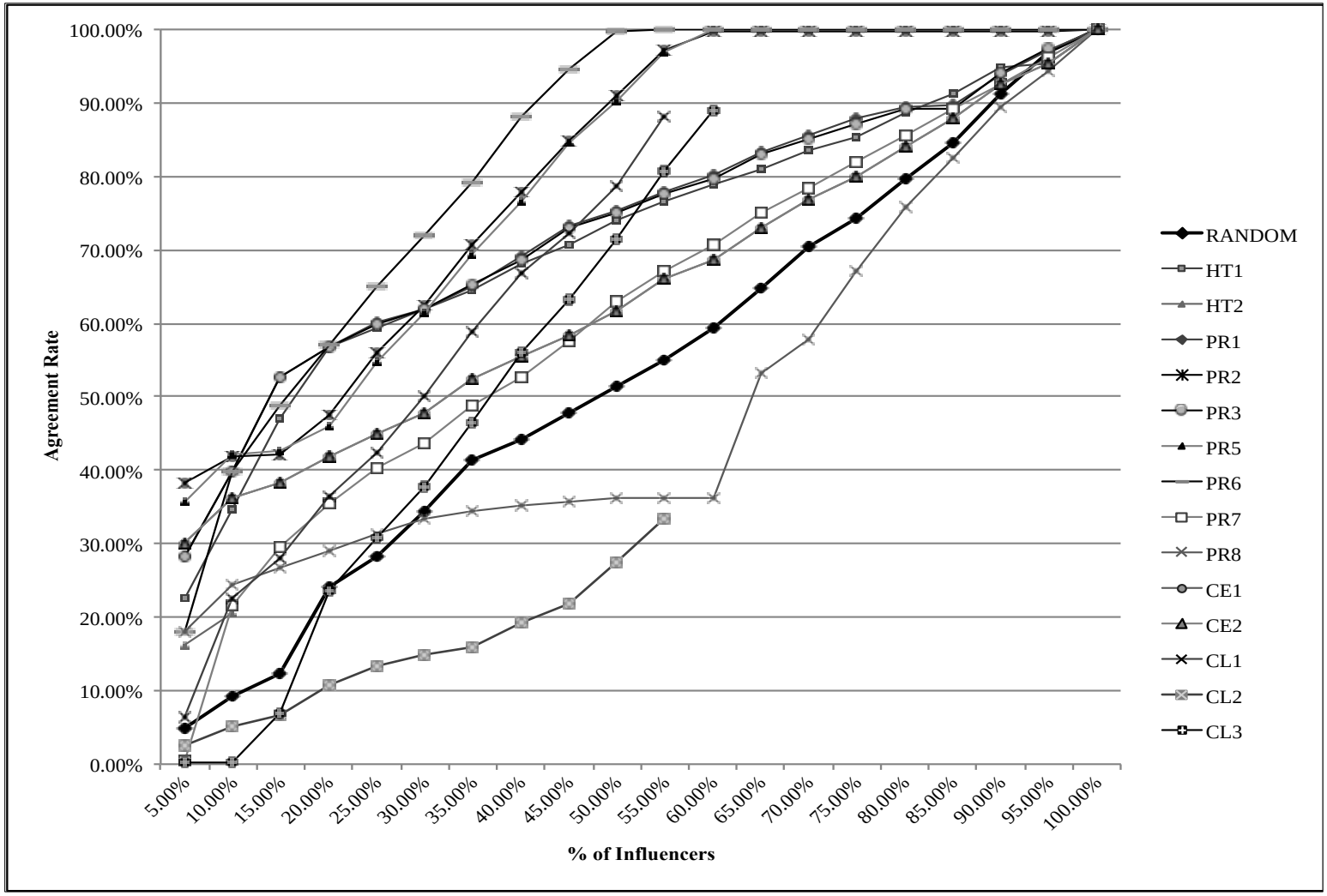
A third metric for evaluating the quality of identified influencers is the agreement rate. The mechanism for capturing the agreement value is that when a covered participant replies to an influencer with positive sentiment, the influencer's agreement value will increase. This can convey how strongly the covered others support an influencer.

Figure 3.12 Agreement Rate for Different Algorithms: Twitter Election Dataset

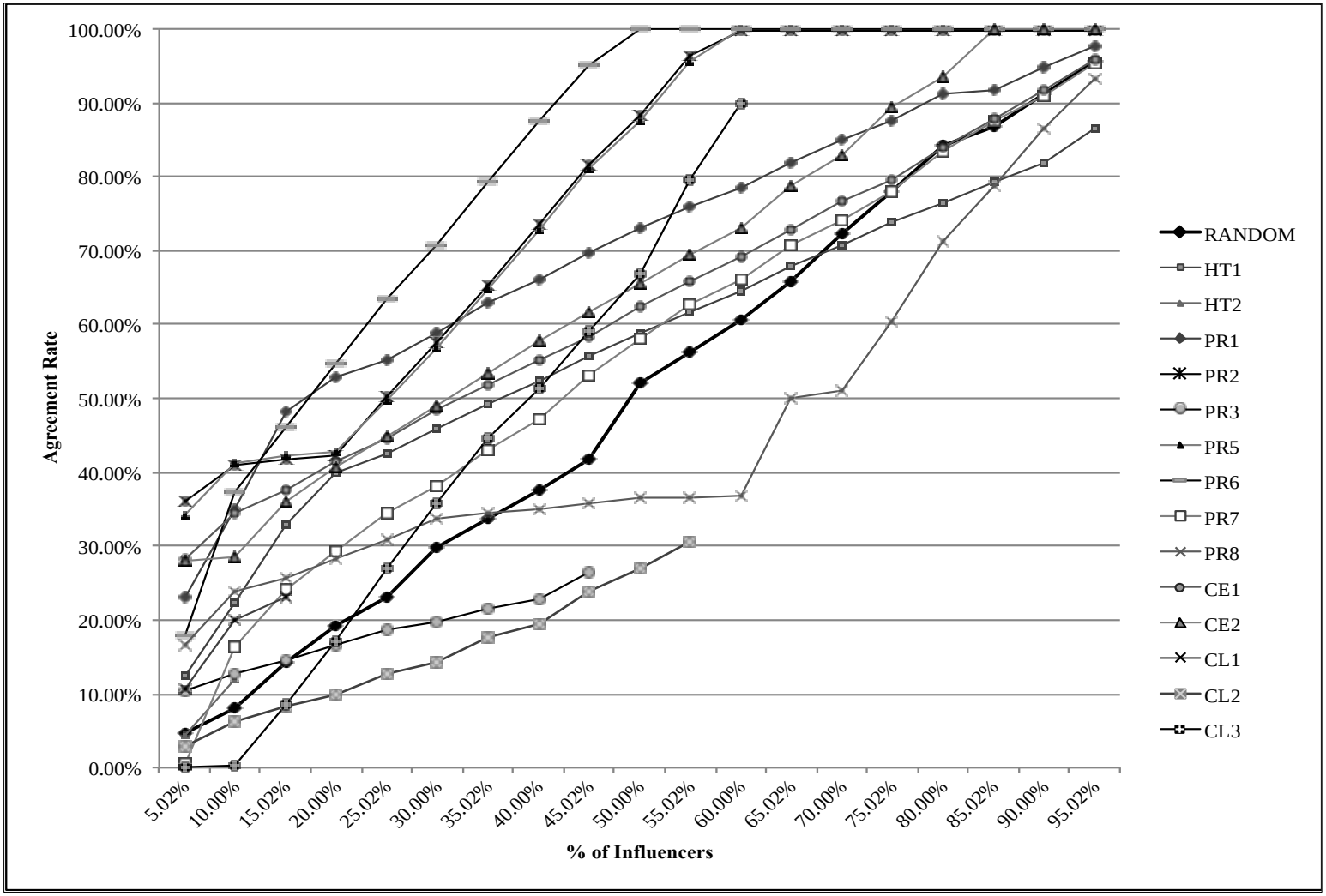
(a) Window 1, *N*=720



(b) Window 2, *N*=1,700



(c) Window 3, *N*=2,510

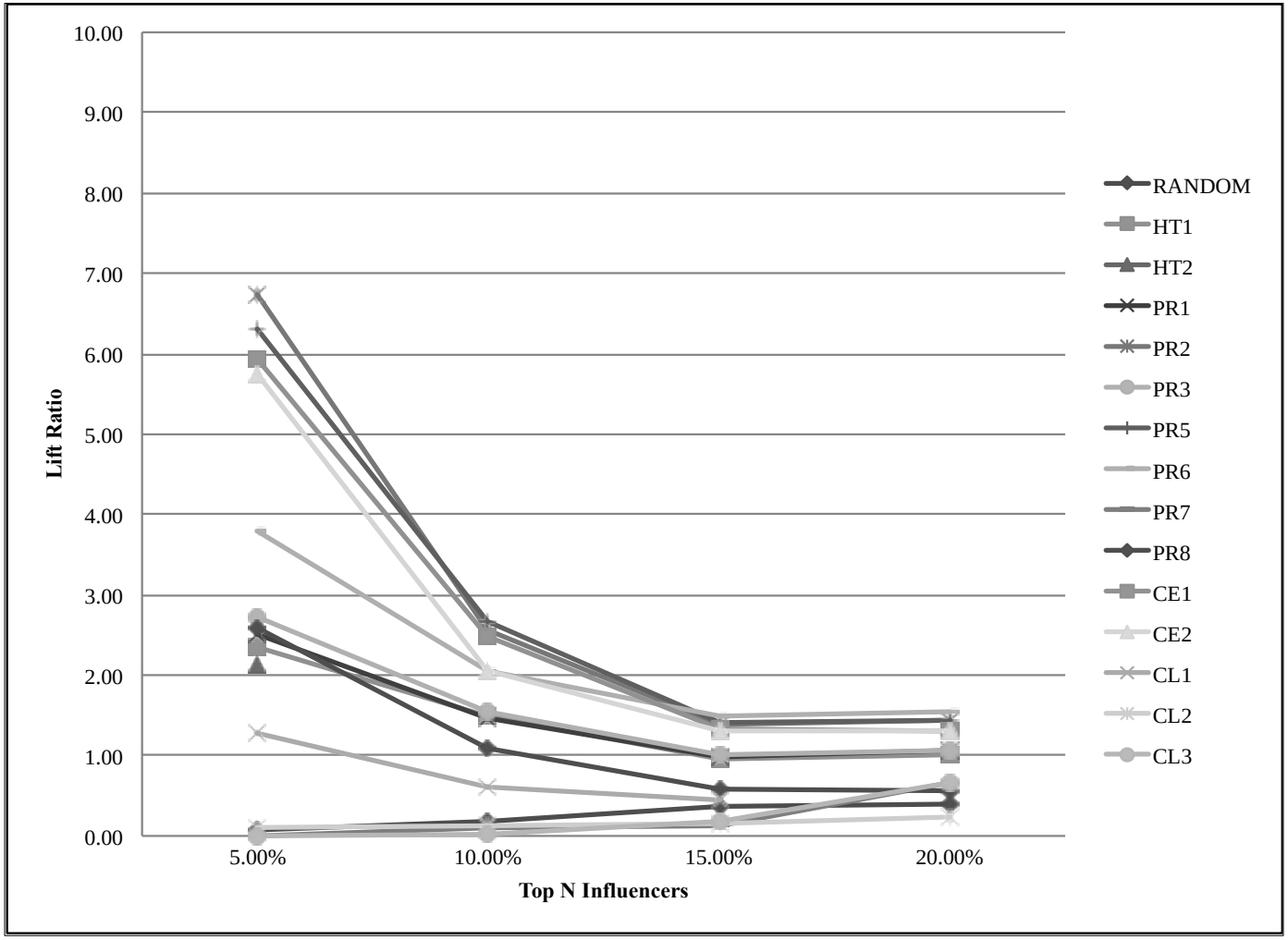


The foregoing results in Figure 3.12 show consistency in the agreement rate metric compared to the results from coverage rate and language diffusion metrics. The influencers identified by PR2, PR5, and PR6 present a higher quality of agreement rate. The influencers identified by PR6 exhibit strong quality in agreement rate, which means not only that the influencers identified by PR6 cover many people, but also that most of these people agree with the influencers' positions. Overall, the results of the agreement rate are fairly similar to those of the coverage rates.

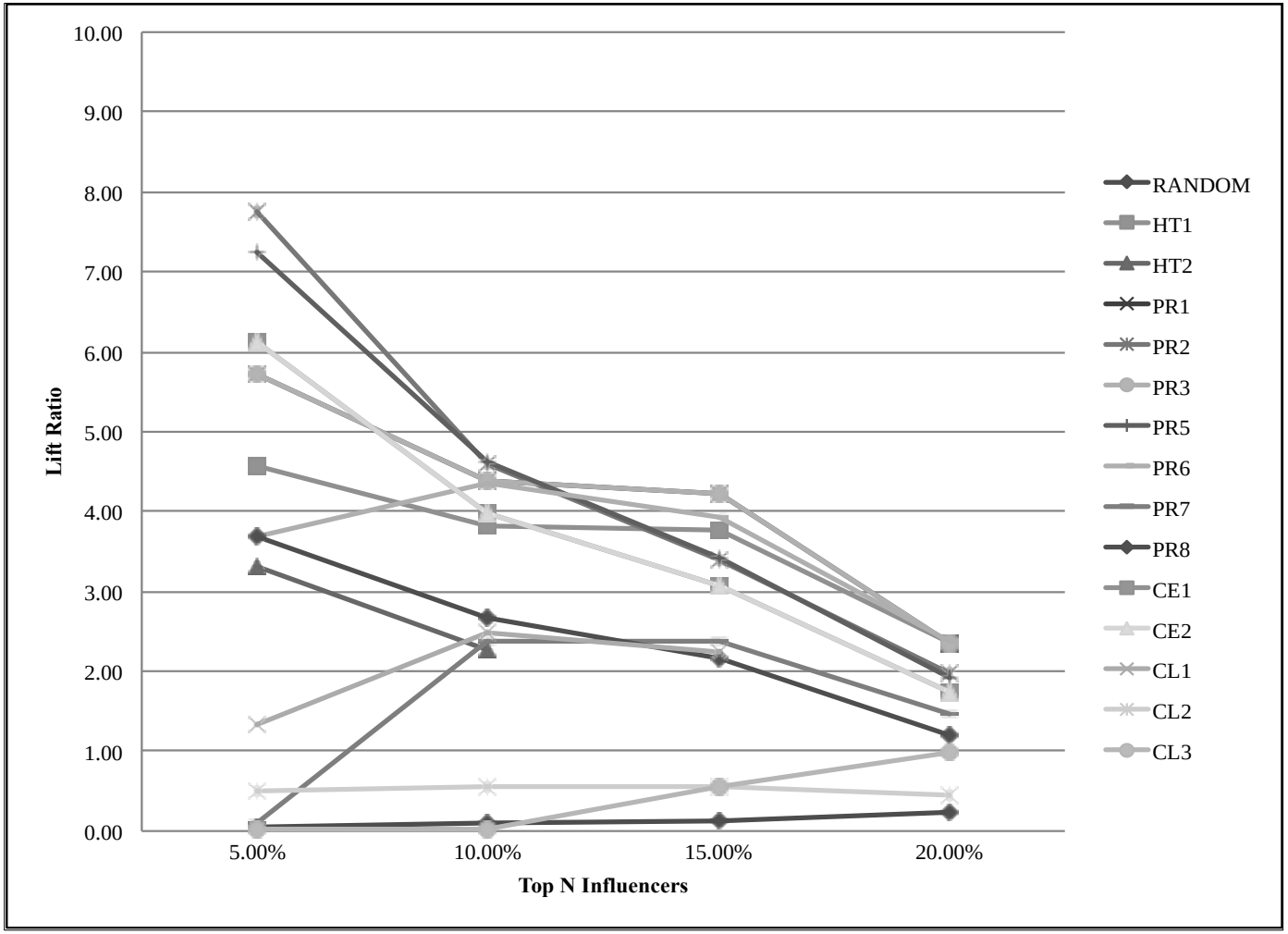
From the three quality metrics, the influencers identified by CL1 and CL3 reveal a steep trending line, similar to the results from PR2, PR5, and PR6. This means that CL1 and CL3 also provide influencers with strong quality across three metrics. However, the influencers identified from PR2, PR5 and PR6 still present a better quality across three different metrics. Figure 3.14 displays lift ratio charts for the top 30% of influencers.

Figure 3.13 Agreement Rate Lift Ratio Charts for Different Algorithms: Twitter Election Dataset

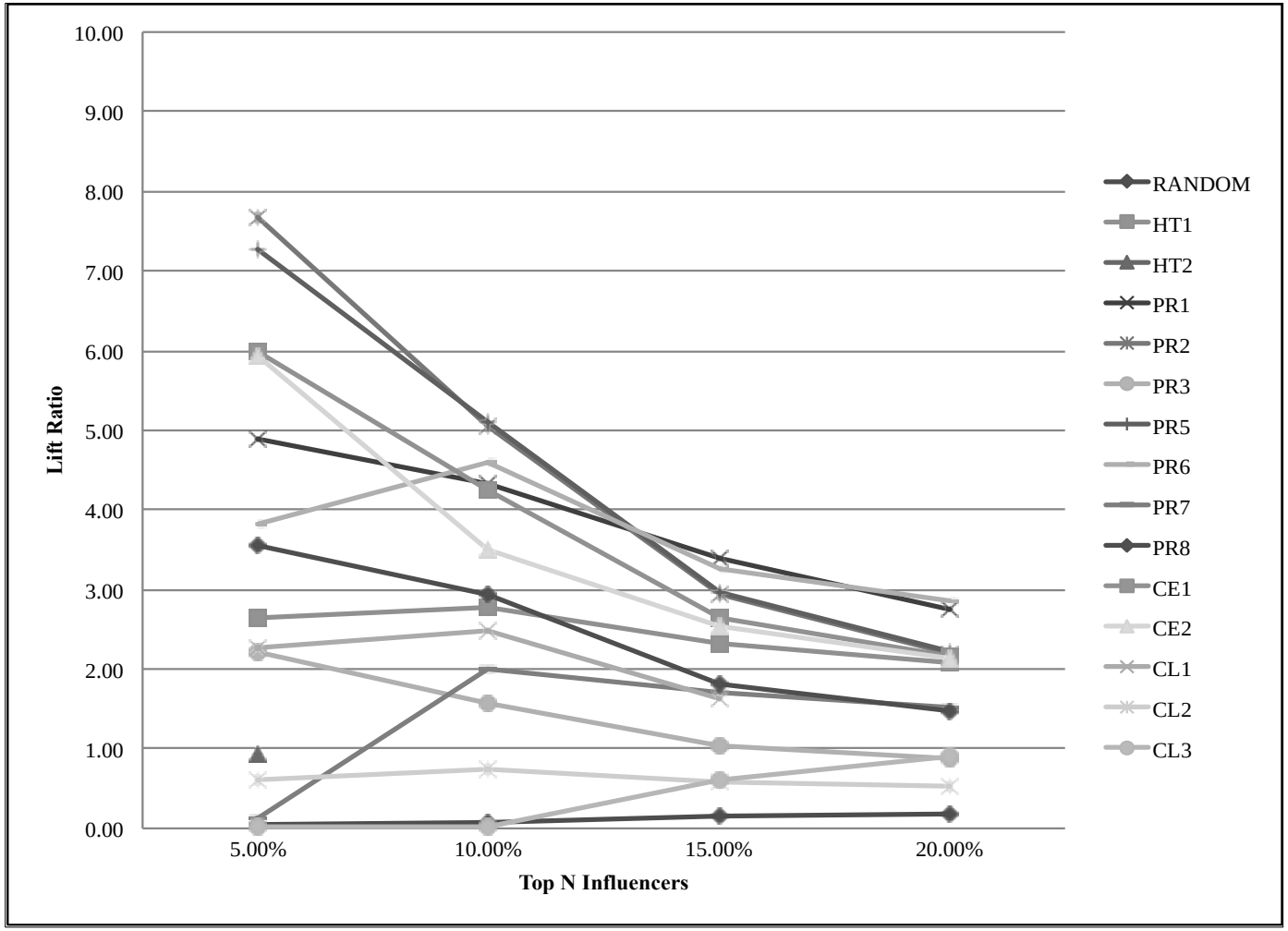
(a) Window 1, *N*=720



(b) Window 2, *N*=1,700



(c) Window 3, *N*=2,510

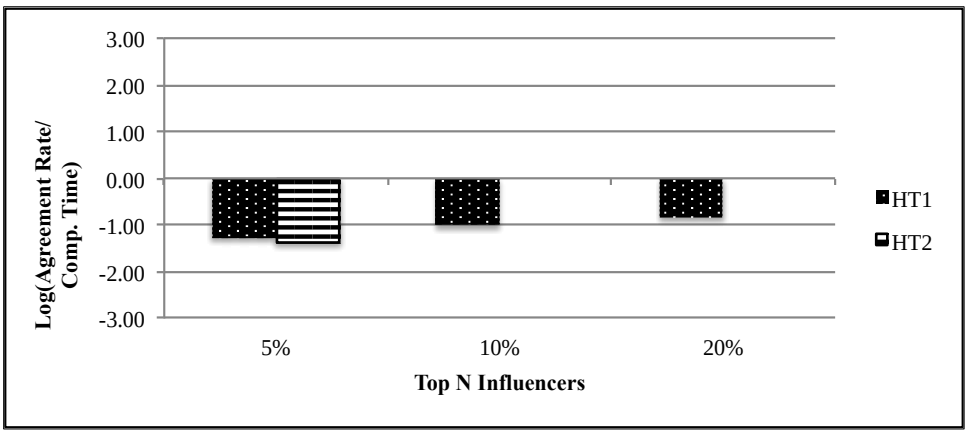


The lift ratio charts in Figure 3.13 show that the influencers identified by PR6 still have the best agreement rate. The influencers identified by PR2 and PR5 also have good agreement rates but not as high quality as the coverage and language diffusion rates. This means that the influencers identified by these algorithms cover many people who are talking about the same subject but hold different opinions. The lift ratio lines also show the same converge trend as those results in converge rate and language diffusion.

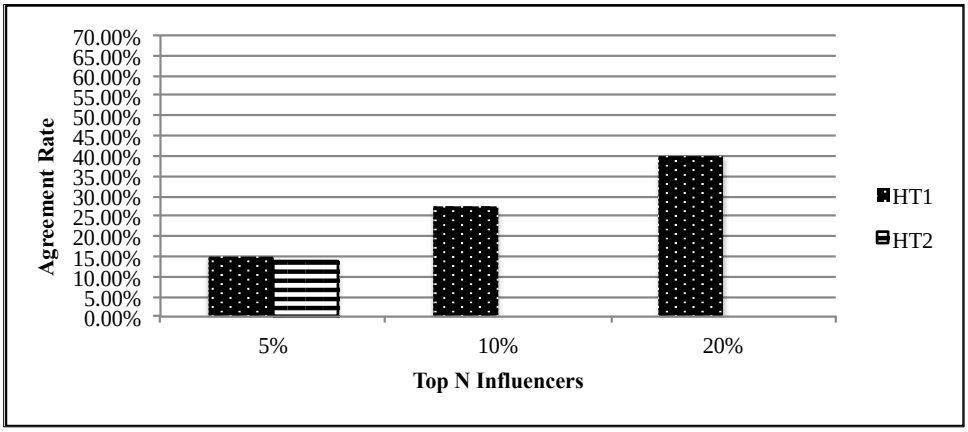
On the other hand, the influencers identified from CE1 and CE2 present a relatively medium to low agreement rate. The results from PR1 and PR3 are somehow unstable. The influencers identified from PR1 and PR3 reveal high agreement rate in window 2 and 3 but not window 1. Overall, the influencers identified from CL2 are relatively low across all three metrics and most of the data windows. Figure 3.14 presents the ratio of the agreement rate to computation time to measure the quality of the identified influencers.

Figure 3.14 Agreement Rate of Top *N*% of Influencers for Different Algorithms in Different Categories: Twitter Election Dataset

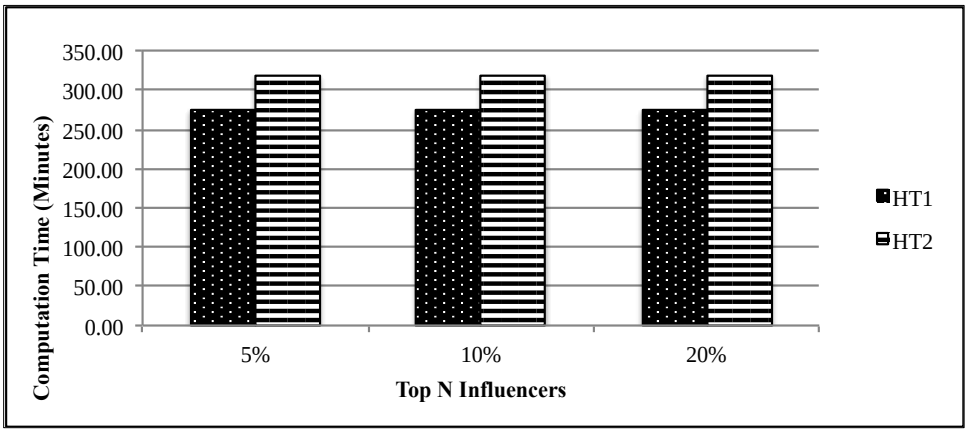
(a-1) HITS-based Algorithms, Window 1, *N*=720



Coverage Rate/ Computation Time

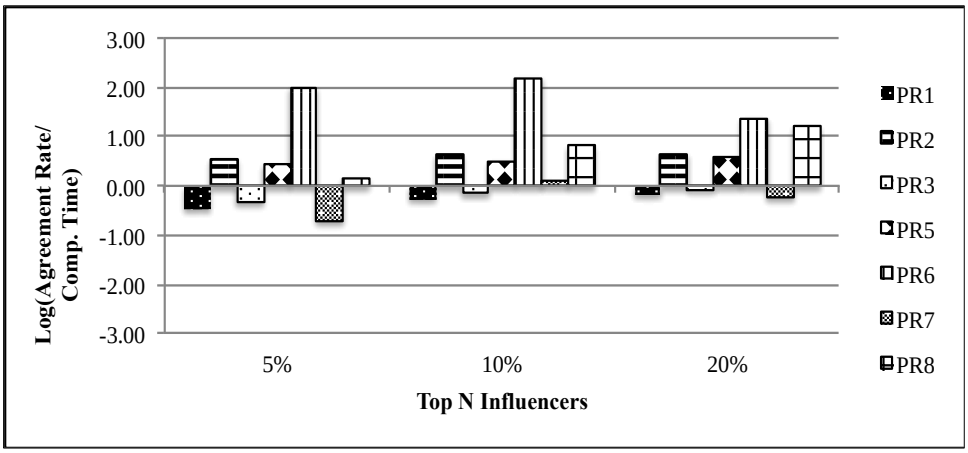


Agreement Rate

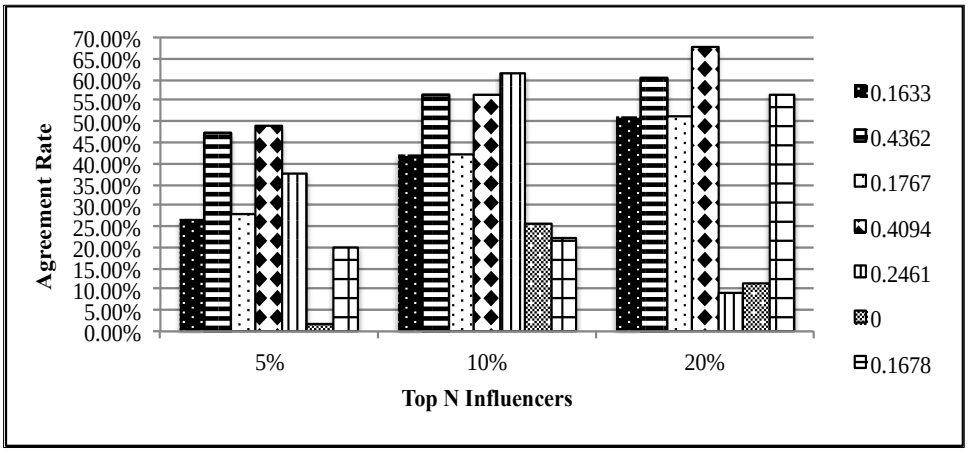


Computation Time

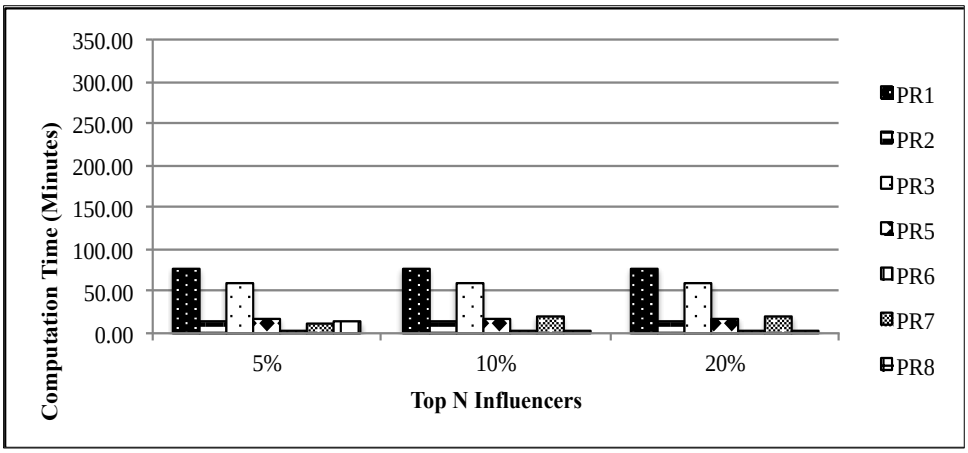
(a-2) PageRank-based Algorithms, Window 1, *N*=720



Coverage Rate/ Computation Time for

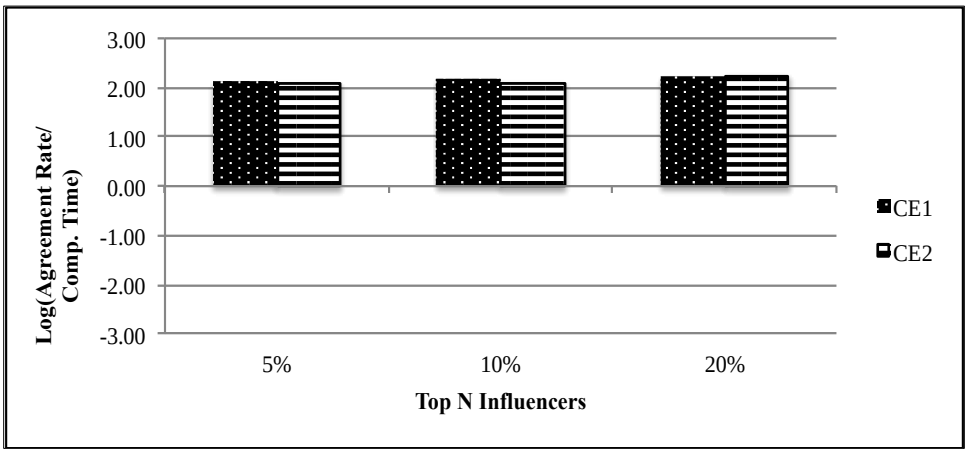


Agreement Rate

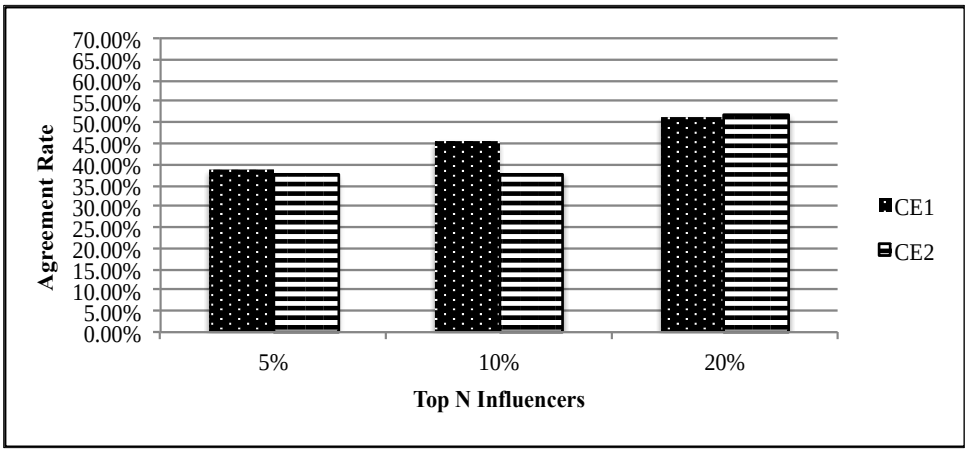


Computation Time

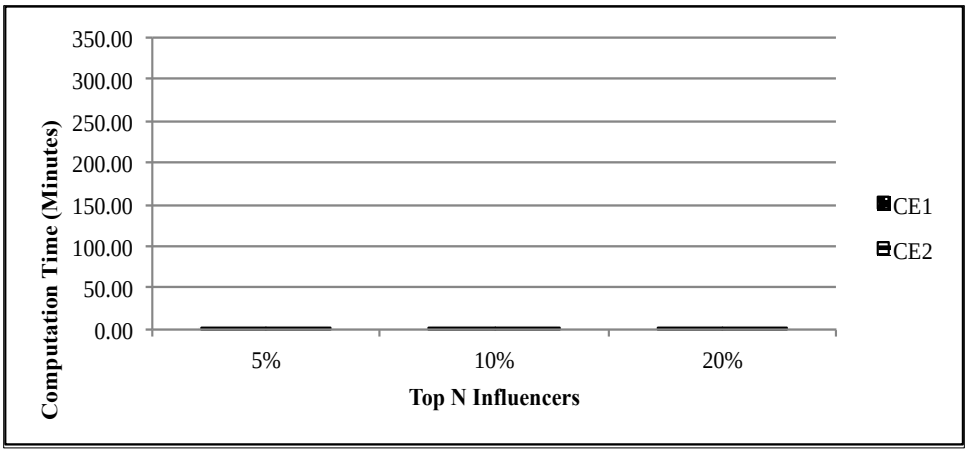
(a-3) Centrality-based Mechanisms, Window 1, *N*=720



Coverage Rate/ Computation Time

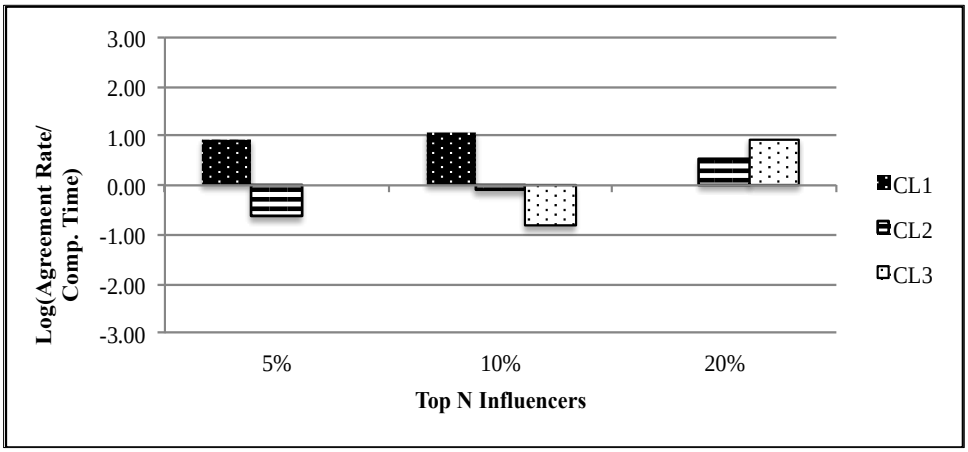


Agreement Rate

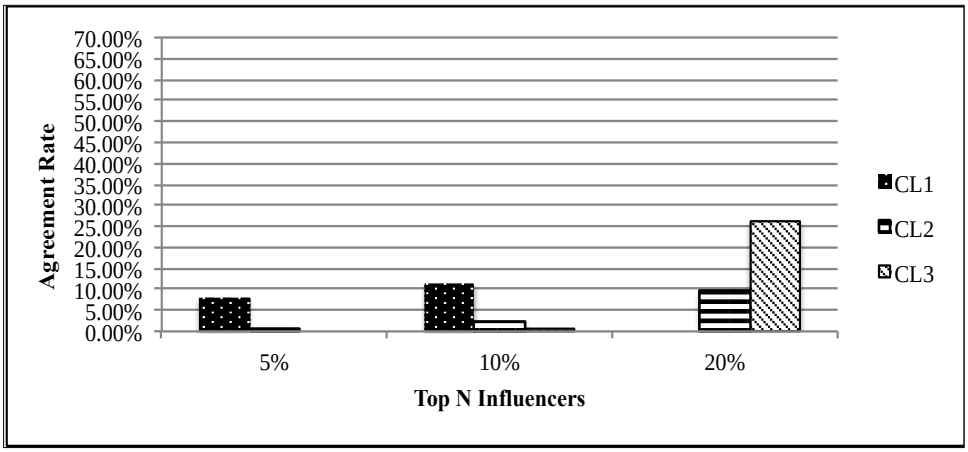


Computation Time

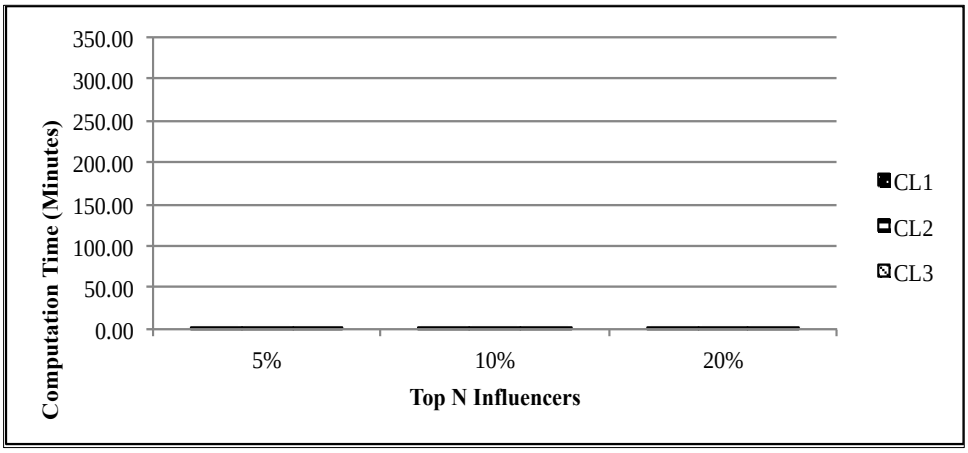
(a-4) Clustering-based Algorithms, Window 1, *N*=720



Coverage Rate/ Computation Time

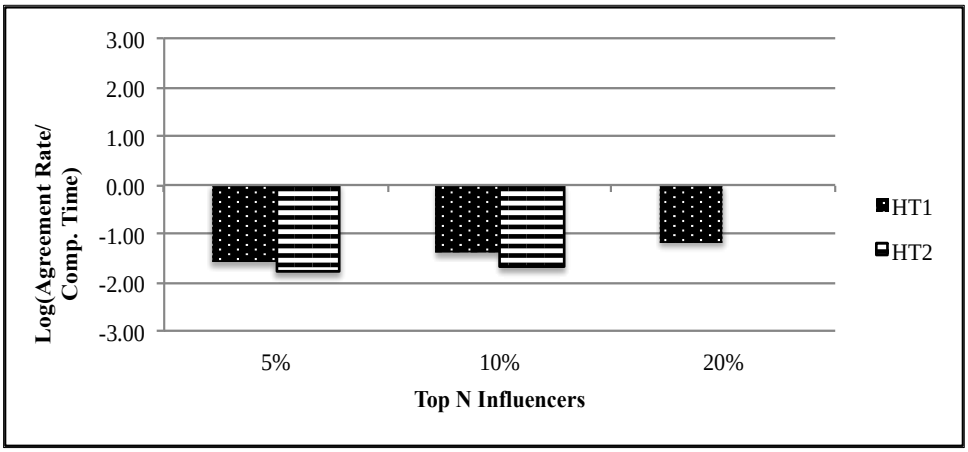


Agreement Rate

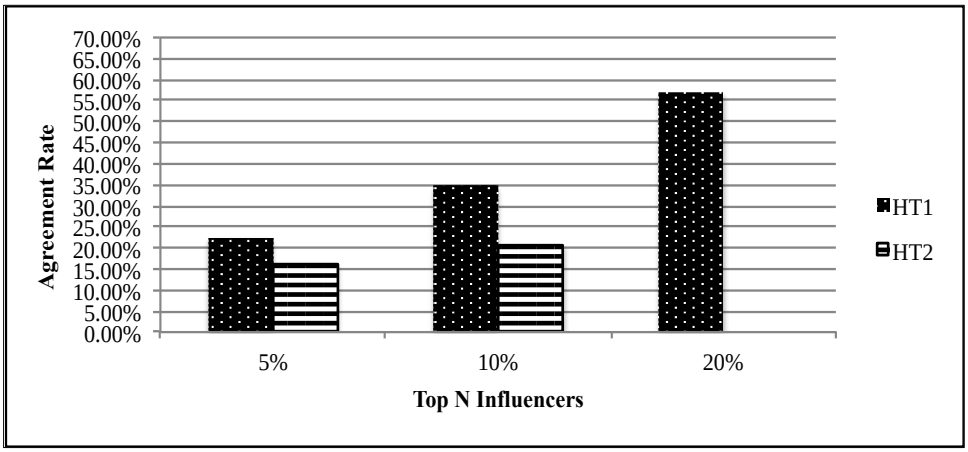


Computation Time

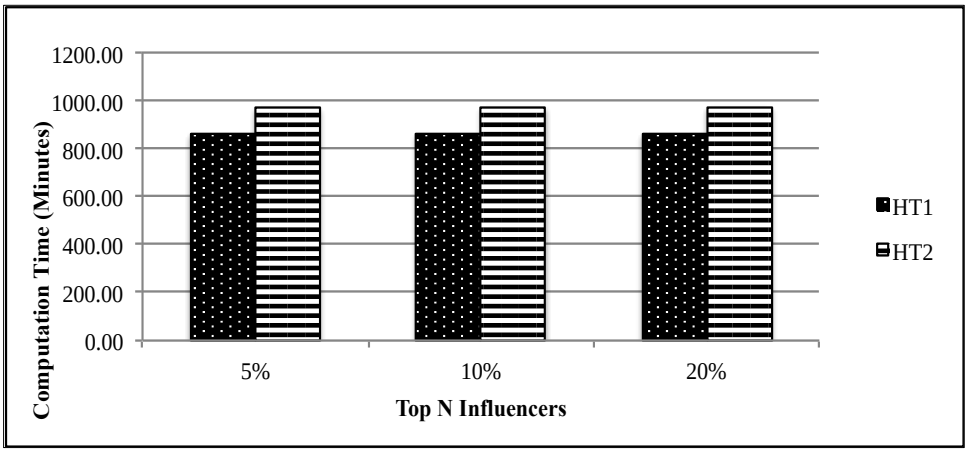
(b-1) HITS-based Algorithms, Window 2, *N*=1,700



Coverage Rate/ Computation Time

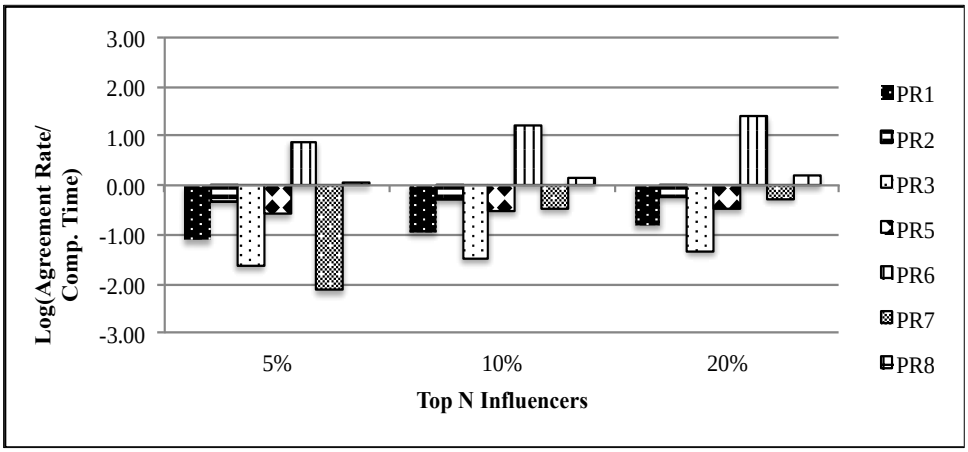


Agreement Rate

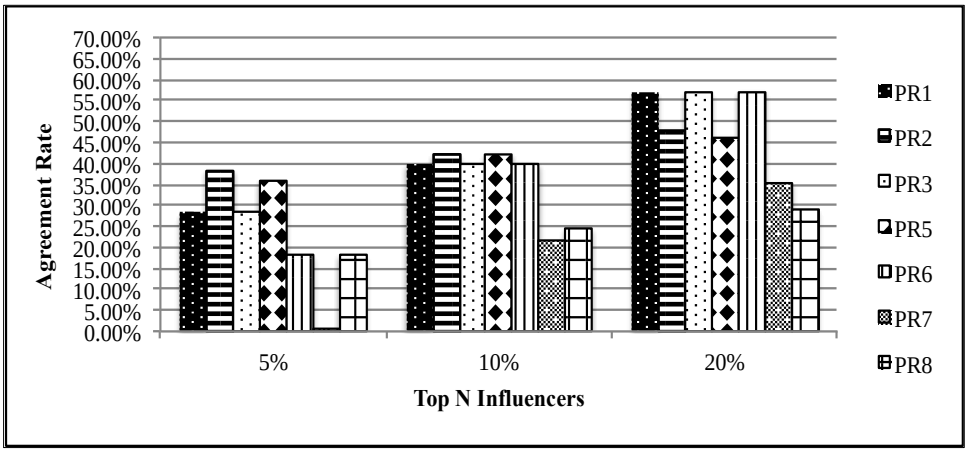


Computation Time

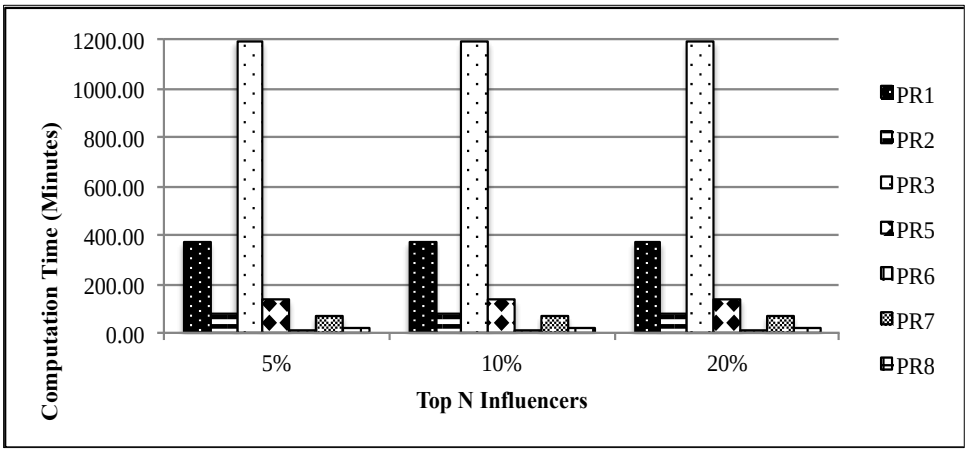
(b-2) PageRank-based Algorithms, Window 2, *N*=1,700



Coverage Rate/ Computation Time

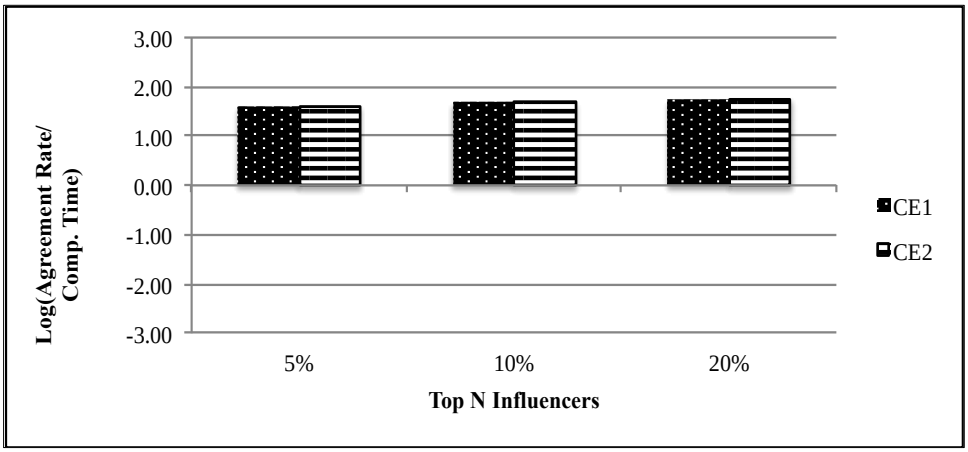


Agreement Rate

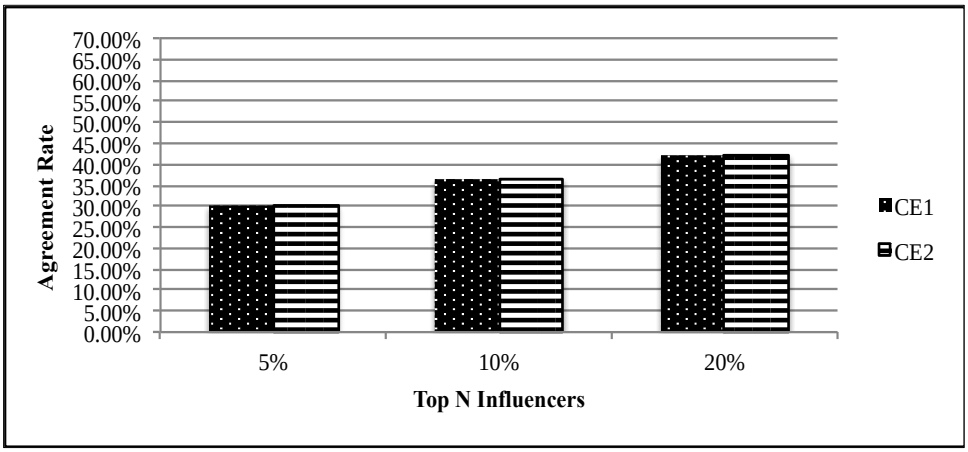


Computation Time

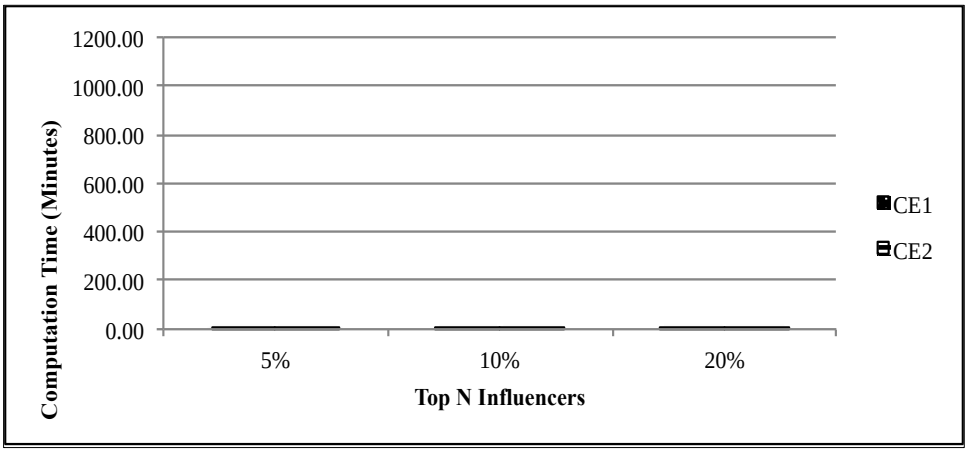
(b-3) Centrality-based Mechanisms, Window 2, *N*=1,700



Coverage Rate/ Computation Time

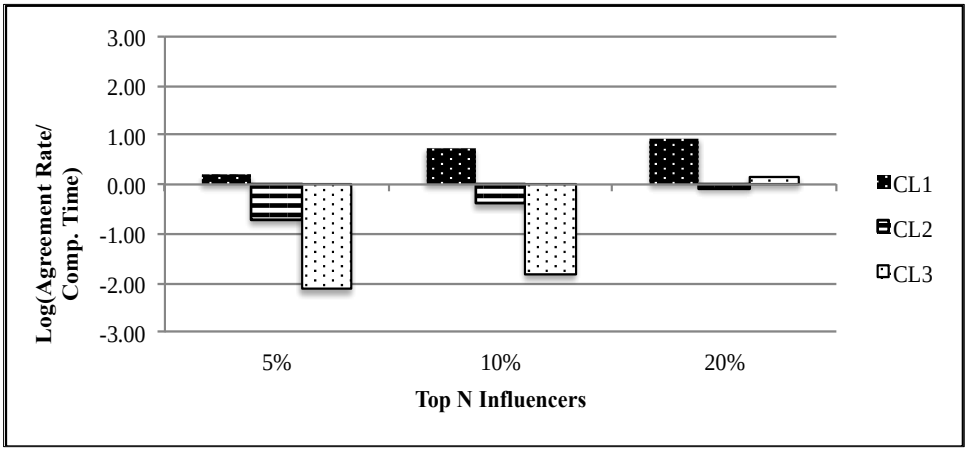


Agreement Rate

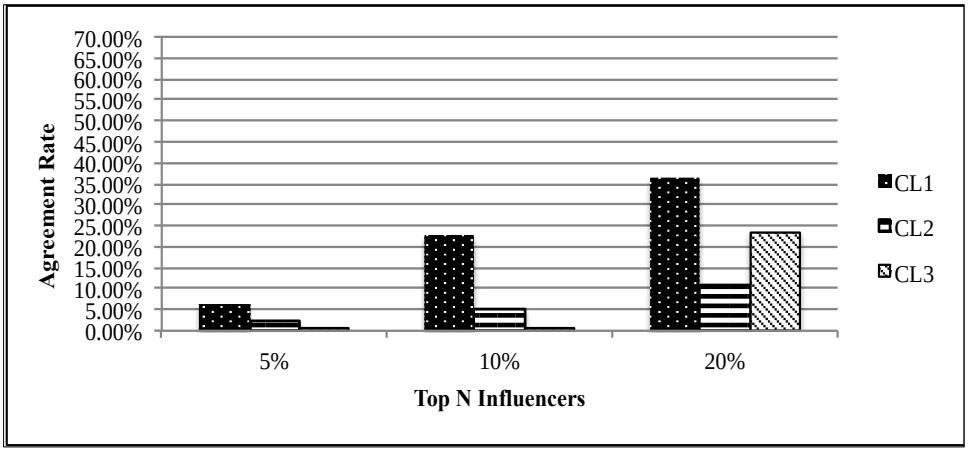


Computation Time

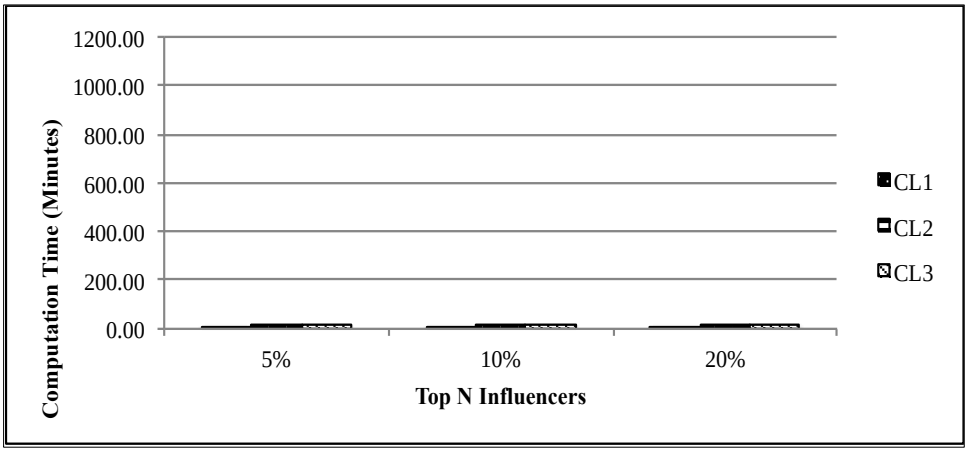
(b-4) Clustering-based Algorithms, Window 2, *N*=1,700



Coverage Rate/ Computation Time

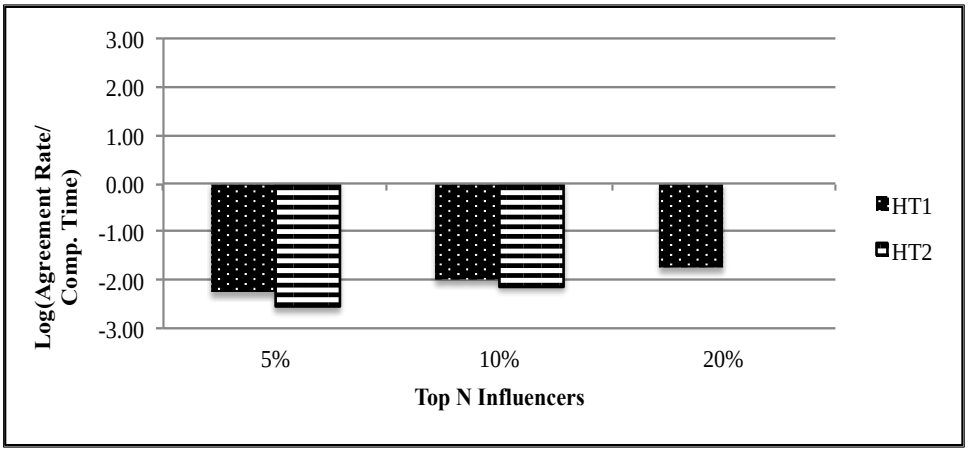


Agreement Rate

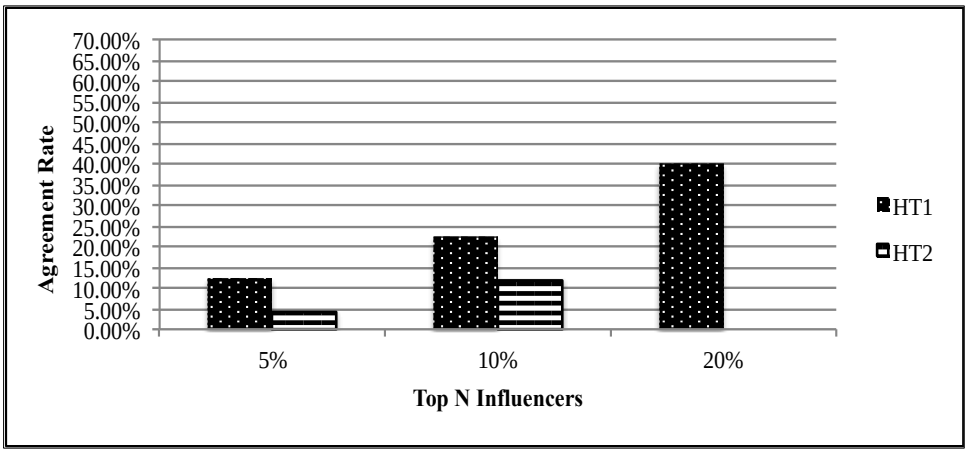


Computation Time

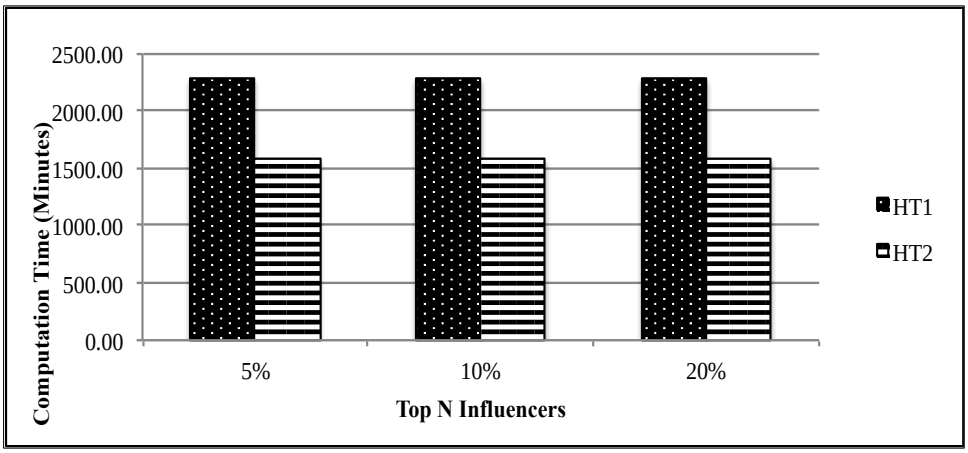
(c-1) HITS-based Algorithms, Window 3, *N*=2,510



Coverage Rate/ Computation Time

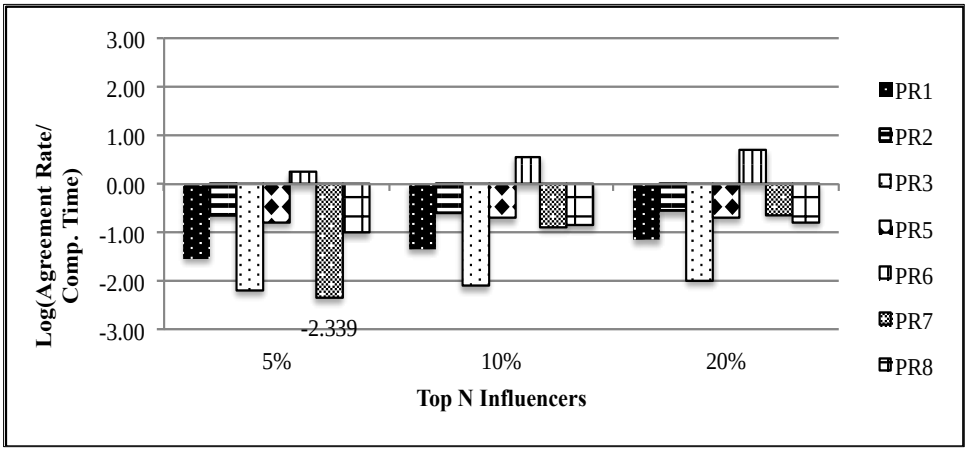


Agreement Rate

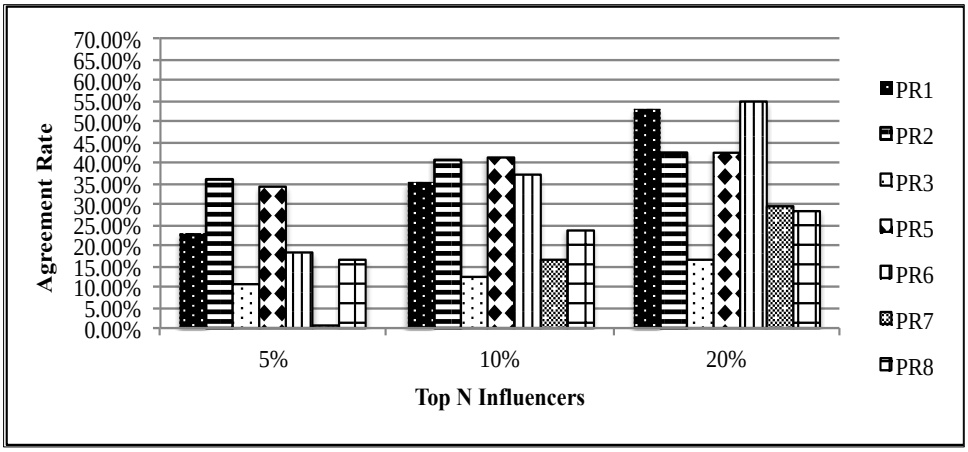


Computation Time

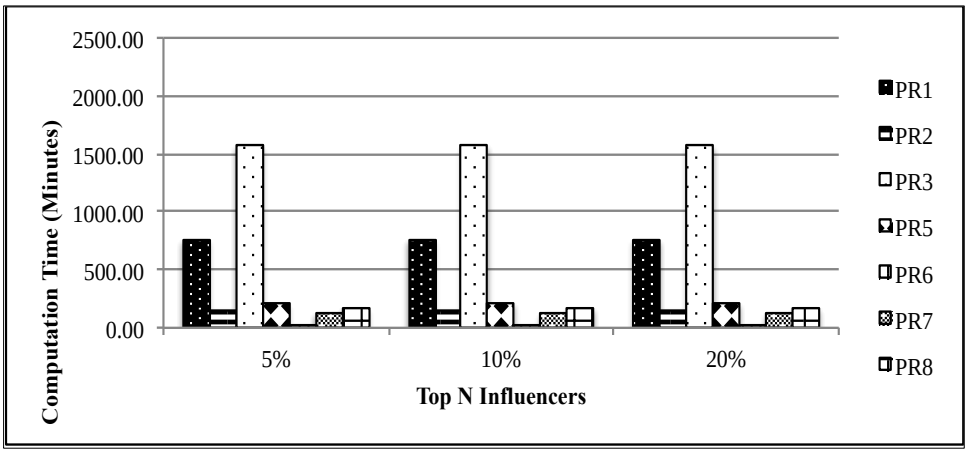
(c-2) PageRank-based Algorithms, Window 3, *N*=2,510



Coverage Rate/ Computation Time

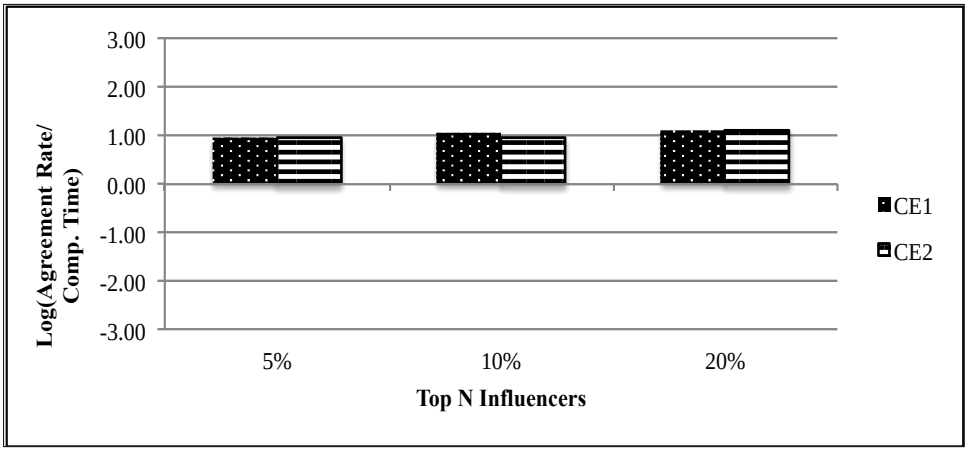


Agreement Rate

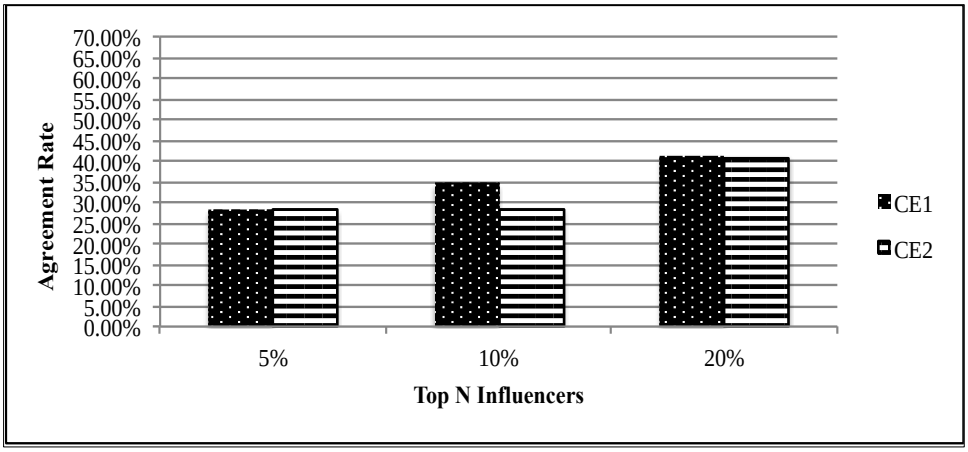


Computation Time

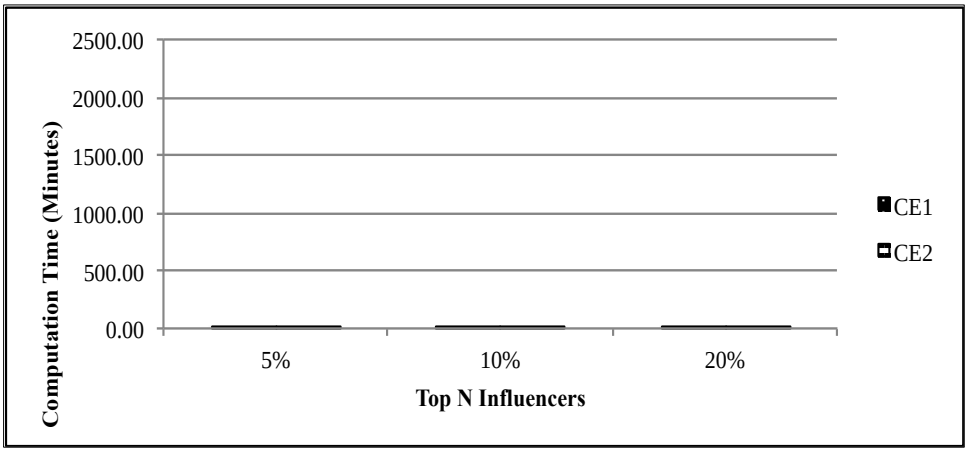
(c-3) Centrality-based Mechanisms, Window 3, *N*=2,510



Coverage Rate/ Computation Time

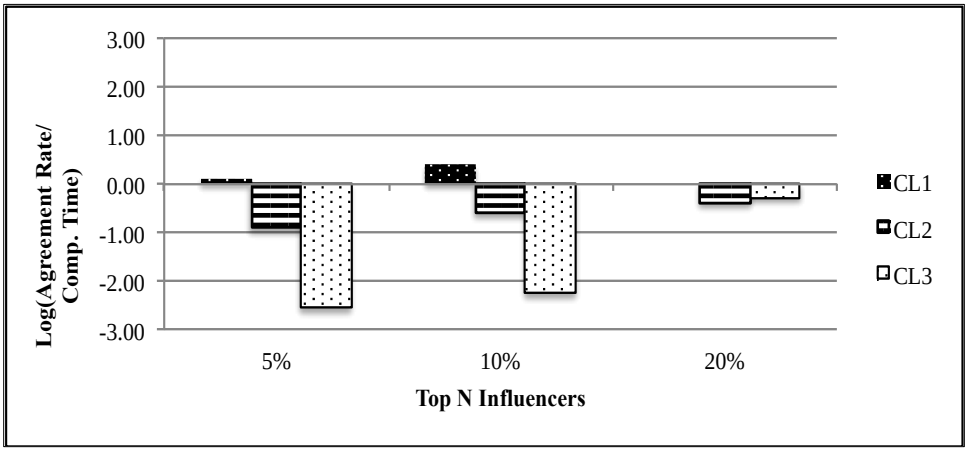


Agreement Rate

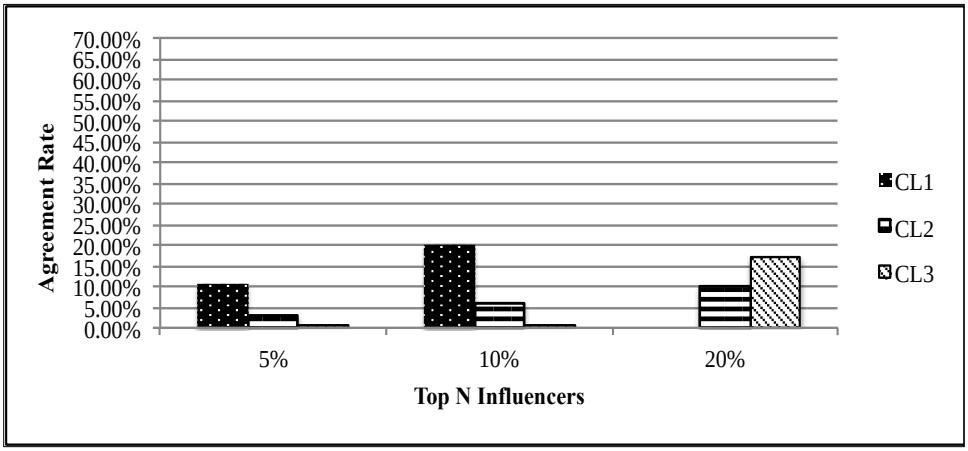


Computation Time

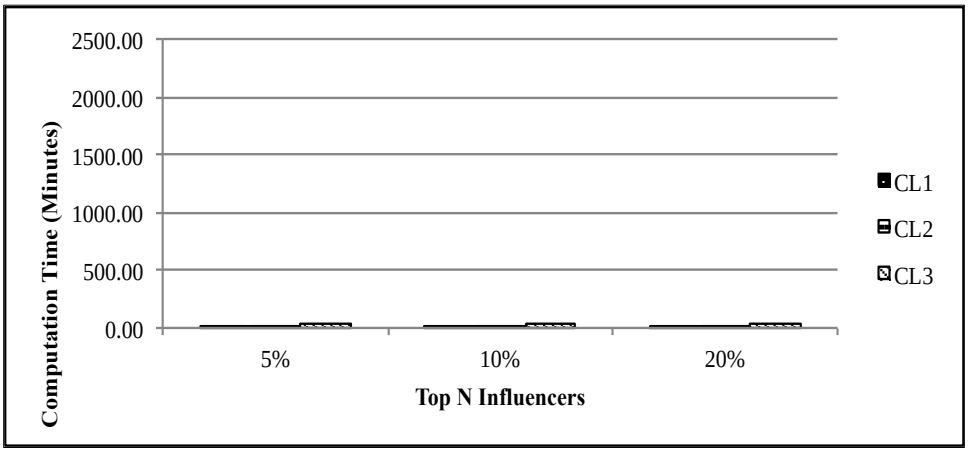
(c-4) Clustering-based Algorithms, Window 3, *N*=2,510



Coverage Rate/ Computation Time



Agreement Rate



Computation Time

The charts in Figure 3.14 show that the influencer identification by PR6, CE1, and CE2 is better than the results from other algorithms in terms of the ratio of the agreement rate to computation time. This result is also consistent with the results from the coverage rate and language diffusion rate metrics. The influencers identified by HT1 are slightly better than those identified by HT2 in terms of their agreement rate. Among the three clustering-based algorithms, CL1 shows better performance in identifying influencers. Based on the results for the three different metrics, this study offers suggestions for ensembles of multiple approaches.

3.8.3. Ensemble Approaches

Based on the results from the multiple experiments, a couple of ensemble approaches were explored to investigate the results of integrating multiple approaches to influencer identification. Ensemble approach refers to the technique of combining multiple mechanisms to produce a single set of results. The ensemble approach produces the results with better quality and lower error than single method (Opitz & Maclin, 1999). For example, the results of the HITS-based ensemble approach produce the influencers identified by both HT1 and HT2. The combined results provide a relatively accurate (intersection) or larger (union) group of influencers than those identified by either HT1 or HT2 algorithm.

First, this investigation analyzes the three different quality metrics for ensemble approaches, based on the different categories of approaches to influencer identification (i.e., HITS-based algorithms, PageRank-based algorithms, centrality-based methods, and clustering-based algorithms). This investigation also selects the best approach from each category based on the ratio of the coverage rate to computation time to form another ensemble approach. The results for the intersections and unions of the influencers identified by these ensemble approaches with respect to the quality metrics follow.

3.8.3.1 Intersection Ensembles

Table 3.4 shows the intersections of top *N* % of influencers identified by different algorithm/methods among each ensemble method and the actual number of identified influencers. This table presents the coverage rate, language diffusion rate, and the agreement rate of the influencers identified from each ensemble method. For example, the first row in Table 3.4 indicates that the intersection of the top 5% influencers identified by the HT1 and HT2 ensemble

includes only one person, with a coverage rate of 0.26%, language diffusion rate of 0.68%, and agreement rate of 0.22%.

Table 3.4 Coverage, Language Diffusion, and Agreement Rates of Influencers in Intersection Ensembles: Twitter Election Dataset

Window 1, *N*=720					
Influencers	**Ensembles**	***N* Influencers**	**Coverage Rate**	**Language Diffusion Rate**	**Agreement Rate**
Top 5%	HITS-based Algorithms	1	0.26%	0.68%	0.22%
	Centrality-based Methods	3	18.53%	0.00%	15.44%
Top 10%	HITS-based Algorithms	2	0.53%	1.36%	0.67%
	Centrality-based Methods	6	26.48%	1.02%	22.37%
Window 2, *N*=1,700					
Influencers	**Ensembles**	***N* Influencers**	**Coverage Rate**	**Language Diffusion Rate**	**Agreement Rate**
Top 1%	HITS-based Algorithms	1	4.81%	0.64%	3.82%
	Centrality-based Methods	16	24.96%	19.56%	20.44%
Top 5%	HITS-based Algorithms	16	10.96%	15.68%	9.98%
	PageRank-based Algorithms	1	4.81%	0.64%	3.82%
	Centrality-based Methods	79	36.50%	37.91%	30.17%
Top 10%	HITS-based Algorithms	43	13.15%	19.14%	12.56%
	PageRank-based Algorithms	2	7.06%	7.71%	5.30%
	Centrality-based Methods	119	36.93%	38.76%	30.67%
	HT1, PR6, CE2, CL1	4	13.84%	0.00%	10.71%
Window 3, *N*=2,510					
Influencers	**Ensembles**	***N* Influencers**	**Coverage Rate**	**Language Diffusion Rate**	**Agreement Rate**
Top 1%	Centrality-based Methods	6	0.76%	0.04%	0.66%
Top 5%	HITS-based Algorithms	17	0.21%	0.78%	0.38%
	Centrality-based Methods	46	5.49%	11.60%	4.24%
	HT2, PR6, CE1, &CL1	3	0.51%	1.78%	0.75%
Top 10%	HITS-based Algorithms	79	8.70%	21.64%	8.38%
	PageRank-based Algorithms	6	0.59%	1.48%	0.94%
	Centrality-based Methods	180	34.73%	43.72%	28.25%
	Clustering-based Algorithms	5	0.00%	0.00%	0.00%
	HT2, PR6, CE1, CL1	18	4.77%	13.69%	4.24%

Table 3.4 shows how the different ensemble approaches identify groups of influencers. Quality metrics of the intersections of the identified influencers are analyzed to see how these ensemble approaches perform in terms of the quality of the identified influencers. Among all the ensemble intersection approaches, the centrality-based methods provide the best quality based on the three metrics. For example, in Window 3, the centrality-based ensemble intersection approach identifies the influencer intersection that provides a 5.49% coverage rate, 11.6% language diffusion rate, and 4.24% agreement rate, which are superior to those from the other ensemble intersection approaches. Ensemble approaches can also be analyzed based on the quality of identified influencers unions.

3.8.3.2 Union Ensembles

Table 3.5 shows the unions of the top *N*% of influencers identified from each ensemble method. For the intersection ensemble method, there is normally fewer influencers identified from different algorithms/methods, but the identified influencers are more convincing than those identified from sign approach. If multiple algorithms/methods in one ensemble method all identify the same group of people as influencers, there is a higher possibility these influencers can make a strong impact.

The union ensemble methods, on the other hand, provide a larger group of influencers than a single approach, which is another option when considering an ensemble method. Suppose the need is to locate a huge group of influencers from the social media network. The union ensembles provide bigger groups of influencers. Table 3.5 presents the influencer results for union ensembles.

Table 3.5 Coverage, Language Diffusion, and Agreement Rates of Influencers in Union Ensembles: Twitter Election Dataset

Window 1, *N*=720					
Influencers	**Ensembles**	**N**	**Coverage Rate**	**Language Diffusion Rate**	**Agreement Rate**
Top 1%	All Algorithms	122	45.54%	50.00%	41.61%
	HITS-based Algorithms	29	11.47%	25.25%	12.30%
	PageRank-based Algorithms	81	43.69%	46.10%	39.15%
	Centrality-based Methods	18	39.28%	36.36%	34.23%
	Clustering-based Algorithms	51	3.18%	6.86%	3.58%

Top 5%	All Algorithms	28	32.48%	13.47%	29.98%
	HITS-based Algorithms	6	5.03%	9.92%	6.49%
	PageRank-based Algorithms	14	28.51%	3.22%	25.28%
	Centrality-based Methods	4	18.53%	0.00%	15.44%
	Clustering-based Algorithms	9	0.44%	1.02%	0.45%
	HT1, PR6, CE2, CL1	126	51.37%	58.39%	49.22%
Top 10%	All Algorithms	53	41.13%	40.93%	36.24%
	HITS-based Algorithms	14	10.33%	25.00%	10.96%
	PageRank-based Algorithms	28	37.33%	31.86%	32.44%
	Centrality-based Methods	8	27.45%	3.98%	23.94%
	Clustering-based Algorithms	21	1.59%	3.64%	1.57%
	HT1, PR6, CE2, CL1	212	56.75%	64.49%	56.15%
Window 2, *N*=1,700					
Influencers	**Ensembles**	**N**	**Coverage Rate**	**Language Diffusion Rate**	**Agreement Rate**
Top 1%	All Algorithms	145	33.40%	36.63%	28.08%
	HITS-based Algorithms	33	22.39%	14.67%	18.47%
	PageRank-based Algorithms	73	29.08%	30.25%	23.77%
	Centrality-based Methods	16	15.50%	23.23%	13.42%
	Clustering-based Algorithms	51	0.96%	0.96%	1.48%
	HT1, PR6, CE2, CL1	55	27.63%	23.82%	23.40%
Top 5%	All Algorithms	601	62.37%	78.36%	56.28%
	HITS-based Algorithms	154	37.47%	34.29%	30.67%
	PageRank-based Algorithms	323	53.39%	71.35%	47.04%
	Centrality-based Methods	93	36.88%	38.49%	30.67%
	Clustering-based Algorithms	250	6.79%	4.15%	8.62%
	HT1, PR6, CE2, CL1	292	45.59%	42.32%	37.32%
Top 10%	All Algorithms	948	75.25%	89.05%	74.01%
	HITS-based Algorithms	297	48.05%	55.93%	41.50%
	PageRank-based Algorithms	587	64.67%	82.14%	63.42%
	Centrality-based Methods	223	40.51%	41.84%	34.36%
	Clustering-based Algorithms	372	8.07%	9.62%	9.11%
	HT1, PR6, CE2, CL1	486	62.43%	67.09%	56.53%
Window 3, *N*=2,510					
Influencers	**Approaches**	**N**	**Coverage Rate**	**Language Diffusion Rate**	**Agreement Rate**
Top 1%	All Algorithms	286	16.39%	24.68%	16.76%
	HITS-based Algorithms	48	1.23%	2.91%	1.60%
	PageRank-based Algorithms	154	12.21%	17.38%	12.62%
	Centrality-based Methods	42	3.97%	5.13%	4.05%
	Clustering-based Algorithms	71	1.56%	3.87%	1.98%

	HT2, PR6, CE1, CL1	88	18.00%	12.43%	14.60%
Top 5%	All Algorithms	912	61.64%	80.40%	62.24%
	HITS-based Algorithms	233	15.50%	30.90%	15.35%
	PageRank-based Algorithms	577	55.09%	69.10%	53.95%
	Centrality-based Methods	576	65.06%	91.61%	61.68%
	Clustering-based Algorithms	346	9.84%	21.99%	11.11%
	HT2, PR6, CE1, CL1	391	39.97%	52.24%	36.44%
Top 10%	All Algorithms	1245	71.31%	90.13%	72.69%
	HITS-based Algorithms	423	32.78%	57.32%	31.73%
	PageRank-based Algorithms	806	65.02%	80.79%	66.67%
	Centrality-based Methods	454	68.10%	97.57%	56.78%
	Clustering-based Algorithms	638	19.98%	47.85%	21.37%
	HT2, PR6, CE1, CL1	665	53.53%	78.70%	51.88%

From the Table 3.5, observe that the "all-algorithm" ensemble approach produces the best quality influencers in most cases. The ensemble approach that integrates the best algorithm from each category also produces good quality influencers. Based on the results for the unions of influencers, this integrated ensemble may be a better choice than the all-algorithm ensemble approach because it is relatively simple.

3.9. Summary

After conducting all of the experiments, computation times for running these different approaches to influencer identification have been presented. Quality of identified influencers, based on three alternative metrics, is measured. Overall, the influencers identified by PR2, PR5, and PR6 yield better quality in terms of the three metrics used in this dissertation. On the other hand, the influencer identification by CE1 and CE2 proved to be relatively efficient in the experiments. The centrality-based methods provide medium-quality influencers, but the computation time is relatively short. Surprisingly, the centrality-based methods analyzed social network relationships without considering text content and user activity and provided an efficient result. Even though the quality of the identified influencers may not be the best, the centrality-based approaches have the most efficient performance.

Based on evaluation metrics used here, results from the experiments demonstrate that alternative approaches perform differently in terms of computation time, quality of identified influencers, and efficiency (ratio of influencer quality and computation time). A social media

analyst decides which approach to adopt depending on demands of the current situation. Generally speaking, when analyzing a relatively small social media network, the computation time will not cause serious impacts in analysis, and those approaches with higher influencer quality should be considered. On the other hand, if the social media network is substantially larger, approaches with lower computation time may be more suitable.

Further, the introduction of ensemble approaches demonstrates that integrating multiple approaches can provide relatively better quality of identified influencers in terms of coverage rate. In recent data analytics development, the computation cost is getting lower and computer capability is getting better. Combining multiple influencer identification approaches allows a company to reach more customers and improve the SMA capability.

Here, results for one of the three cases recognized earlier have been presented and analyzed. The same experimental procedure has been conducted for the other two cases – March Madness and Kentucky Derby. Findings for these are presented in Appendices III and IV, respectively. Overall, the results from the March Madness and Kentucky Derby datasets are consistent with those from the Twitter Election dataset. From the results of March Madness datasets, it shows that PR2, PR5, and PR6 also provide the best quality of influencers across three different metrics. The influencers identified via CE1 and CE2 also yield strong quality across three different metrics. However, the computation time of PR6 algorithm is as low as those of CE1 and CE2, but the quality of identified influencers is better than those from CE1 and CE2. PR6 becomes the better selection in terms of the bang-to-buck ratio quality across three different metrics in the Twitter March Madness Datasets. On the other hand, the quantity of identified influencers is relatively low in the Kentucky Derby Twitter datasets. The main difference in the Kentucky Derby datasets is that both HT1 and CL2 provide the strong quality of influencers across three different metrics. It turns out that HT1 and CL2 also preform pretty well in the bang-to-buck ratio across three different metrics. This can be concluded that HT1 and CL2 are more appropriate to be applied to the relatively small datasets.

In sum, this study reviews literature related to social media, social influencer, and influencer identification to build up a relatively comprehensive understanding of this SMA implementation. Based on this review, this study designs multiple experiments to implement the influencer identification approaches in multiple social media networks. The assessment of these social influencer identification approaches gives the guidance for academicians to develop future

SMA research and offer practitioners one example of SMA implementation for decision support. For example, marketing researchers and practitioners can apply this implementation to improve marketing strategy by employing influencer identification. The identified influencers can produce strong word-of-mouth (WOM) and diffuse information effectively. The experiment procedures and results in this study also present a relatively comprehensive theoretical foundation for understanding SMA implementations for influencer identification. Future works can be based on this study to apply influencer identification in different scenario.

Chapter 4. Conclusion

4.1. Contributions

This dissertation examines the current literature related to social media analytics (SMA) and develops an integrated, unifying definition of business SMA, thus providing a nuanced starting point for future business SMA research. This definition gives practitioners a relatively clear understanding when designing, developing, and evaluating their own SMA initiatives. It also benefits educators by providing an intellectual base for conveying the knowledge of business SMA and introducing it to more people. This dissertation identifies several benefits of business SMA and elaborates on some of them while presenting recent empirical evidence in support of the argument in this study. This helps practitioners understand how SMA can provide assistance to their organization. . To help organizations be better informed about investing in SMA initiatives, this dissertation provides an example that illustrates the application of SMA to extract valuable information from big data in support of decision-making. The dissertation also describes several challenges facing Business SMA today, along with supporting evidence from the literature, some of which also offer mitigating solutions in particular contexts. The main purpose of documenting these challenges is to alert researchers to future directions for investigation. These unsolved problems need to be emphasized for future development of this area.

Another contribution in this dissertation is the introduction of a framework of SMA-based decision-making. This framework leads SMA researchers in the direction of adopting a decision support point of view. Based on varying business needs, SMA can support manager in a relevant decisional phase of a business process. For example, the Intelligence stage allows a company to ferret out customer opinions. It helps marketing strategy development and also customer relationship management. The problem recognition and opportunity detection features support new product development process to design a more customer-oriented model. At the same time, this framework can be applied in business analytics and intelligence training to give a relatively comprehensive view of SMA in decision support field.

Growing social media usage, accompanied by explosive growth in SMA, has resulted in increasing interest in finding automated ways of discovering social influencers (i.e., opinion leaders) in online social interactions. Yet, there has, heretofore, been no extensive study investigating the relative efficacy of all current methods in specific settings. This dissertation investigates and reports on the relative performance of multiple methods on Twitter datasets

containing between them tens of thousands to hundreds of thousands of tweets. This dissertation furthers the research area of social influencer identification from Social Media. Researchers can use this dissertation as a reference to extend social influencer identification to a next level. For example, one can apply the identification approach with lowest computation time to identify influencers from a huge, but different, network in a different setting (i.e., internal social media for a multinational corporation). This dissertation also provides practitioners with a roadmap when adopting influencer identification approaches to his/her own company, =deciding which approach is more suitable in specific business settings.

4.2. Limitations

One limitation of this dissertation is its focus on very recent literature. Social Media Analytics related discussions could be expanded using a larger time window for literature extraction and review. Here, the literature search focuses only on "social media analytics/intelligence" as keywords. It may be worthwhile expanding the search to also include papers using "social network analysis," "sentiment analysis," "text mining," and "web mining" as key words. While these keywords may net several irrelevant papers, insofar as the focus is on analyzing only social media content, this dissertation may yet avoid overlooking important work. This dissertation has also considered a few conference proceedings papers and industry white papers for insights not available, as yet, in the form of published academic journal articles. The framework of SMA-based decision-making is a conceptual framework with no empirical evidence to support. More empirical work could be included to support the function of each component in this framework.

Further, this dissertation adopts multiple influencer identification approaches into the experiment design. However, there is, so far, no objective metric to evaluate the quality of the result of identified social influencers. The metrics adopted in this dissertation can only explain the "quality" of the results in specific ways (e.g., coverage rate). This dissertation collects data only from Twitter. Because the nature of networks in each social media is different, the results are limited to the networks in Twitter. To provide a relatively precise conclusion, this dissertation's collection of data from Twitter is based on different events and data windows to improve the quality of experiment results. However, only one event is discussed in details. The virtual machine used in the experiments is not advanced enough to execute experiments for much larger datasets. Thus, the results of this dissertation could be limited by the data size. Given the infancy

of academic SMA research, future research based on this dissertation will address several of the cited limitations.

4.3. Future Works

First, the literature review part of this dissertation can be extended to more academia papers and/or industrial white papers. More empirical evidences will be included to understanding benefits and challenges of applying SMA in the business domain. More detail will be incorporated in the framework of SMA-based decision making to provide a sharper picture for application to practice and education.

Second, assessment of social influencer identification approaches will include different datasets from different social media. Forum data will be included to execute the same experiments as another comparison group. This may provide confirmatory results to support the conclusions in this dissertation. At the same time, more advanced information technology (e.g., MapReduce) will be adopted to analyze big datasets within shorter time periods.

A relatively objective integrated metric should be designed to evaluate the quality of identified influencers. Multiple criteria including text feature, network structure, and user behavior will be considered simultaneously when measuring the quality of influencers. Integrated metrics will help to provide a relatively comprehensive measurement of the quality of identified influencers. Time variables will be also integrated into metrics to put both time and quality into consideration.

Appendices

Appendix I: Influencer Identification Approaches

1. HITS-based Algorithms

HT1: The study done by Jing & Lizhen (2014) designs a modified HITS algorithm, namely, HITS_FEATURE algorithm to identify the influencers from a large Chinese microblog site, Sina. This research defines a directed network G = (V, E) based on the comment behavior between users. First of all, this research adapts the Analytic Hierarchy Process (AHP) approach, which requires experts to evaluate the relative importance of each factors related to evaluate influencers. In here, these factors are defined from user activities such as the number of posts, followers, and replies, and are evaluated by experts to build up the AHP matrix.

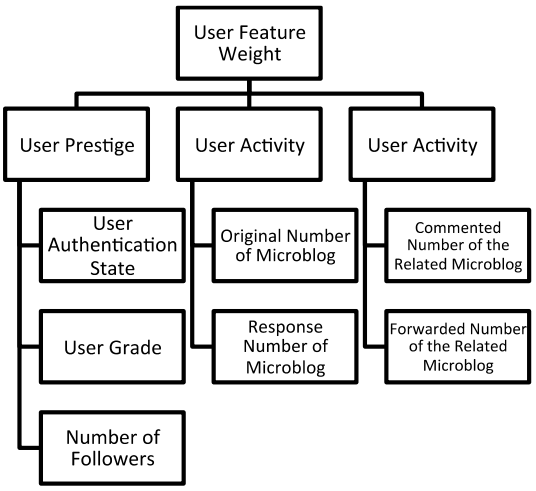


Diagram of the Hierarchical Structure (Jing & Lizhen, 2014)

A comprehensive score of each user is calculated based on this AHP matrix to represent the *feature weight* of the user.

$$W = Prestige + Activity + Influence$$

Secondly, HITS_FEATURE algorithm deploys sentiment analysis to measure the sentiment orientation of each comment. The authors defines the *sentiment weight* as following:

$$\boldsymbol{e_{ij}} = \begin{cases} \boldsymbol{n_{pos} + n_{neg}} \\ \boldsymbol{0} \end{cases}$$

Where $\boldsymbol{e_{ij}}$ represents the sentiment weight of tie (edge) from user *i* to user *j*, n_{pos} is the number of positive comments from user i to user j, and n_{neg} is the number of negative ones. If user i does not comment on user *j*'s post, the value of $\boldsymbol{e_{ij}}$ is 0.

After receiving the feature weight of each user and the *sentiment weight* of each tie, the authority weight and hub weight of each user are analyzed by the HITS_FEATURE algorithm through the iterative process. The value of authority weight and hub weight of user *j* in *k*-th recursion is describe as following:

Authority Weight:

$$\boldsymbol{a(i)_k} = \sum_{\boldsymbol{j:j \to i}} \boldsymbol{h(i)_{k-1}} \times \boldsymbol{w_j} \times \boldsymbol{e_{ji}}$$

Hub Weight:

$$\boldsymbol{h(i)_k} = \sum_{\boldsymbol{j:i \to j}} \boldsymbol{a(i)_{k-1}} \times \boldsymbol{w_j} \times \boldsymbol{e_{ij}}$$

where $\boldsymbol{a(i)_k}$ is the value of authority weight of user *i* in the *k*-th step, $\boldsymbol{h(i)_k}$ is its value of hub weight, $\boldsymbol{w_i}$ is its feature weight, and $\boldsymbol{e_{ij}}$ is its sentiment weight. After the HITS_FEATURE algorithm achieves the convergence, the nodes with highest value of authority weight are recognized as the influencers. Li et al. (2013) then compare the HITS_FEATURE algorithm with the original HITS algorithm by evaluating the quality of the results of identified influencers. Depend on artificial rating, the authors claim that HITS_FEATURE algorithm improve the original HITS algorithm in the quality of influencer identifications.

HT2: Li et al. (2013) modified the original HITS algorithm to identify influencers from Twitter network. The authors define a directed network G = (V, E) based on comment behavior, which V is the set of nodes representing users, and E is the set of ties (edges) representing the comments relationship between user *i* and user *j*. If user *j* comments on user *i*'s tweet, there is a directed tie from *j* to *i* in the network G. Firstly, the author s define two categories of factors: Professional Competence and Value of Expression. Professional Competence is the factors relevant to user

activities (e.g., the number of tweets, followers, replies, etc.) and Value of Expression is the factors based on sentiment score of each tweet (e.g., number of positive words). Next, this study put these two categories of factors into Supporting Vector Machine (SVM) to generate candidate influencers. However, the authors argue that filtering out the negative will improve the performance of influencer identification. Thus, this study applies sentiment analysis to all the posts and divides all the comments into supportive (positive comments) and opposite (negative comments). Only supportive ties are counted into the algorithm, and the opposite ties (so called *Pest*) are ignored. Compared to the baseline linkage-based HITS algorithm, this study argues that the HITS_PEST algorithm provides a better quality in influencer identification.

2. PageRank-based Algorithms

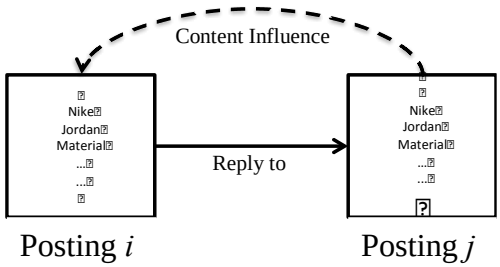
PR1: Unstructured text data are one main product of social media usage. One main stream of SMA researches is to focus only on text data and apply PageRank algorithm to identify influencers. The pre-processing process employs text mining, sentiment analysis, topic modeling to quantify these text contents, and these values become another part of input variables combing with linkage structure to execute PageRank algorithm for influencer identification. Zhou et al. (2009) design a OpinionRank algorithm based on a comment network $G = (V, E, W)$, where G is a directed network, V represents the users as nodes, E represents the edges based on commenting another node's posts, and W represents the opinion scores associated with the edges. The authors apply sentiment analysis to measure the value of the opinion orientation of each comment, normalizing to the values between +1 and -1. These values serve as the strength of ties (edges) between nodes. If user i posts comment on user j's post with an opinion score of +1, the sentiment polarity of the edge from user i to user j is positive with the degrees of positivity of 1. Combining opinion scores of each edge and linkage structure, this study run the OpinionRank algorithm to identify influencer from an Epinions dataset. The OpinionRank algorithm is as follows:

$$\boldsymbol{OR(i) = (1 - d) + d \sum_{j \in B_i} \frac{PR(j) * w_{ji}}{N_j}}$$

where $\boldsymbol{N_j}$ is the total number of out-degree of node j, $\boldsymbol{OR\ (j)}$ is the OpinionRank value of node j, and $\boldsymbol{w_{ji}}$ is the opinion score from i to j. After comparing with the original PageRank algorithm results, the authors content that the OpinionRank algorithm improves the quality of influencer

identification.

PR2: Cheng et al. (2012) design a IS_Rank algorithm to identify influencer. In spite of using sentiment to measure the weight of linkage between users, this research includes the idea of content influence between users. This study argues that when post *i* replies to post *j* and uses the same words as post *j*, the author of post *i* receives content influence from the author of post *j*.



Content Influence from Post *j* to Post *i*

This article first define the BBS network G = (*V*, *E*), where *G* is a directed network, *V* represents the users as nodes, *E* represents the edges based on commenting another node's posts. The authors then assign the edge weights by calculating the influence power between any two users. The total influence *H* between L_1 and L_2 is:

$$\boldsymbol{H_{L_1,L_2} = T_{L_1,L_2} * I_{L_1,L_2}}$$

where the emotional influence $\boldsymbol{T_{L_1,L_2}}$ is the average sentimental score L_1 receives from L_2, the content influence $\boldsymbol{I_{L_1,L_2}}$ is the average number of overlap words between L_1 and L_2, and the influence power $\boldsymbol{H_{L_1,L_2}}$ is the product of the emotional influence and the content influence.

After receiving the total influence weight of each edge, the IS_Rank is modified from the original PageRank as follows:

$$\boldsymbol{IR(i) = (1 - d) + d \sum_{j \in B_i} \frac{IR(j)}{H_{j,i}}}$$

where $\boldsymbol{IR\ (j)}$ is the IS_Rank value of node *j*, and $\boldsymbol{H_{ji}}$ is the total influence score between *j* and *i*. This study recruits four students to manually evaluate the result of identified influencers and

argues that this IS_Rank algorithm can effectively identify the influencer.

PR3: Xiao & Xia (2010) design a LeaderRank algorithm to identify influencers from a Bulletin Board System (BBS) network based on a comment network *G* = (*V*, *E*, *W*, *C*), where *G* is a directed network, *V* represents the users as nodes, *E* represents the edges based on commenting another node's posts, *W* represents the opinion scores associated with the edges, and C is the belonged community of each user. This study emphasizes that influencer should be identified within the interest group rather than among all users. Hence, this study firstly applies topic modeling to cluster users to different group based on their posts. Afterward, the authors execute the LeaderRank algorithm to identify influencer in each group. The LeaderRank algorithm is as follows:

$$\boldsymbol{LR(i) = (1 - d) + d \sum_{j \in B_i} \frac{LR(j) * w_{ji}}{N_j}}$$

where $\boldsymbol{N_j}$ is the total number of out-degree of node *j*, $\boldsymbol{LR\ (j)}$ is the OpinionRank value of node *j*, and $\boldsymbol{w_{ji}}$ is the opinion score from *i* to *j*. After comparing with the different approaches, the authors argue that LeaderRank algorithm performs more efficiently than other approaches.

PR4: Jiang et al. (2013) apply sentiment analysis to measure the value of calculate the link weights between nodes. Differently, their research proposes an improved PageRank building on the Hadoop MapReduce environment to improve the performance of influencer identification. "MapReduce is a programming model for processing and generating large dataset." (Dean & Ghemawat, 2010, p. 72) The *map* function generates a set of key/ value pair and assigns to *reduce* function located in multiple machines for parallel processing. The Hadoop MapReduce framework takes care of parallelization to achieve better performance the running in single process/ machine (Dittrich & Quiané-Ruiz, 2012). The MapReduceRank algorithm is as follows:

$$\boldsymbol{MR(i) = (1 - d) + d \sum_{j \in B_i} \frac{MR(j) * w_{ji}}{N_j}}$$

where $\boldsymbol{N_j}$ is the total number of out-degree of node *j*, $\boldsymbol{MR\ (j)}$ is the MapReduceRank value of node *j*, and $\boldsymbol{w_{ji}}$ is the opinion score from *i* to *j*. This study applies the MapReduceRank system on a Chinese online forum called Tianya Club and argues that the accuracy rate to receive the same

group of influencers as the list provided by the official Tianya Club is higher than the original PageRank algorithm. Meanwhile, the computation time is 7.5 times faster than the original PageRank algorithm in their experiment.

PR5: Ziyi et al. (2013) simply calculate the similarity score of text contents and sentiment preference to measure the integrated influence power between nodes and apply the integrated influence score to PageRank algorithm to identify influencer

$$\boldsymbol{PR(i) = (1 - d) + d \sum_{j \in B_i} \frac{PR(j)}{U_{j,i}}}$$

where $\boldsymbol{PR\ (j)}$ is the PageRank value of node *j*, and $\boldsymbol{U_{ji}}$ is the integrated influence score between *j* and *i*. The integrated influencer score is the product of sentiment influence and content influencer. . In their Sina BBS experiment, they use the coverage ratio to measure the performance of different algorithms. Among six algorithms, their approach reaches the highest coverage ratio, which represents the ability of the influencers in influencing the other nodes.

PR6: Zhai et al. (2008) investigate the replying activities in social media to propose their interest-field based algorithm, FieldPR algorithm, and compare it with other influencer identification approaches. In this study, the authors define a BBS network G = (*V*, *E*, *W*), where *G* is a directed network, *V* represents the users as nodes, *E* represents the edges based on replying another node's posts, and *W* represents the weight of edge, which is measured by the number of receiving replies and of its followers.

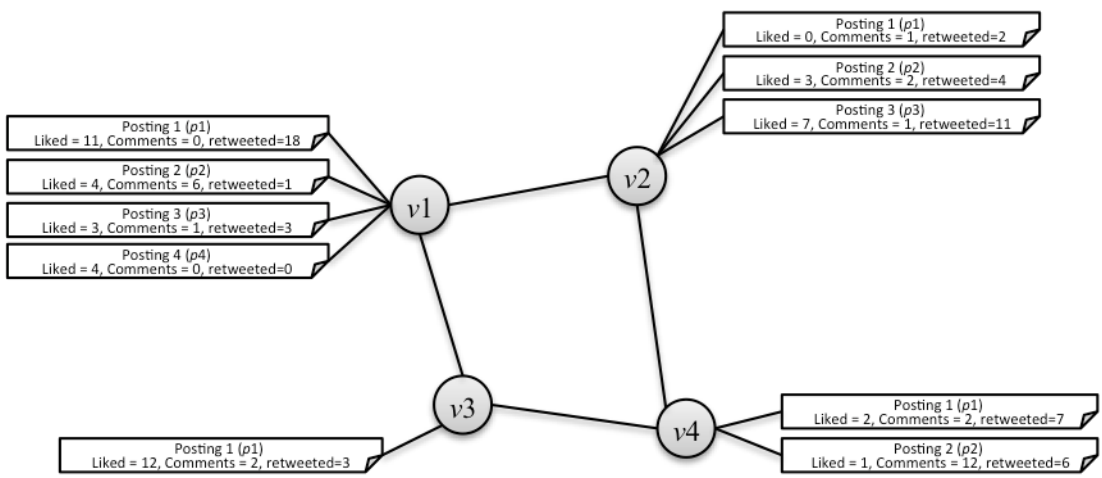
$$\boldsymbol{FieldPR(i) = (1 - d) + d \sum_{j \in B_i} \frac{FieldPR(j) * w_{ji}}{N_j}}$$

where $\boldsymbol{N_j}$ is the total number of out-degree of node *j*, $\boldsymbol{FieldPR(j)}$ is the *FieldPR* value of node *j*, and $\boldsymbol{w_{ji}}$ is the weight from *i* to *j*. In this study, the authors identify influencer from interest groups. To cluster users (nodes) to different interest groups, they use two different approaches: Board-based and Article-chain. These two different approaches are called FieldPR_Board and FieldPR_ChainCluster algorithms. The FieldPR_Board algorithm first clusters nodes into different groups based on the board topics. On the other hand, the FieldPR_ChainCluster clusters

nodes based on the topic among the article-chain. If two users participate the same article-chain discussing same topic, these two users will be clustered into the same interest group.

PR7: Hajian & White (2011) apply activity variables to design an InfluenceRank algorithm. The authors first define a social network graph G = *(V, P, E)*, where *V* is the node representing each user, *P* is the post of each node *V*, and *E* is the edge between nodes. In this study, there are couple social media activities are considered: $\boldsymbol{F(v)}$ is the number of followers of user $\boldsymbol{v}$, $\boldsymbol{P(v)}$ is the number of posts of node $\boldsymbol{v}$, $\boldsymbol{L(p)}$ is the number of "like" received in each post $\boldsymbol{p}$, $\boldsymbol{C(p)}$ is the number of comments received in each posts $\boldsymbol{p}$, and $\boldsymbol{RT(p)}$ is the number of propagations (retweets) of each post $\boldsymbol{p}$ (Figure 10.) $\boldsymbol{LCRT}\ (\boldsymbol{v},\boldsymbol{p})$ is a function that determinates the number a user $\boldsymbol{v}$ has commented, liked, or propagated (retweeted) on a particular post $\boldsymbol{p}$ in a network.

$$\boldsymbol{LCRT}\ (\boldsymbol{v},\boldsymbol{p}) = \begin{cases} \mathbf{1} & \boldsymbol{if}(\boldsymbol{v}\epsilon\boldsymbol{L}(\boldsymbol{p}) \cup \boldsymbol{C}(\boldsymbol{p}) \cup \boldsymbol{RT}(\boldsymbol{p})) \\ \mathbf{0} & \boldsymbol{Otherwise} \end{cases}$$



Network Structure of InfluenceRank Algorithm

Second, the authors develop multiple factors to model the influence power of each node *V*:

1. Popularity ($\boldsymbol{\delta(v)}$): A non-linear function in the range of [0, 1] using the ratio of the followers of a user $\boldsymbol{v}$ to the maximum followers a network indicates the popularity of user $\boldsymbol{v}$.

$$\boldsymbol{\delta(v) = \frac{\ln(F(v) - \min_{v' \in V} F(v'))}{\ln(\max_{v'' \in V} F(v'') - \min_{v' \in V} F(v'))}}$$

2. Ratio of Affection ($\boldsymbol{ROA(v,p)}$): The proportion of the number of followers who has commented on, liked, or propagated (retweeted) a post $\boldsymbol{p}$ of user $\boldsymbol{v}$. This measures the rate of influence power from a user $\boldsymbol{v}$ to their followers.

$$\boldsymbol{ROA(v,p) = \frac{\sum_{v' \epsilon F(v)} LCRT(v',p)}{F(v)}}$$

3. Magnitude of Influence ($\boldsymbol{MOI}$): The root mean square of $\boldsymbol{ROA}$ of all the posts by the user $\boldsymbol{v}$. This indicates the influence power made by user $\boldsymbol{v}$ in a network.

$$\boldsymbol{MOI(v) = \sqrt{\frac{\sum_{p' \epsilon P(v)} (ROA(v,p'))^2}{P(v)}}}$$

The original PageRank algorithm measures the importance of nodes based on the linkage structure in a network. Using above factors, PageRank algorithm is modifies to the InfluenceRank algorithm for improving the accuracy of influencer identification by adding activity variables into consideration. The InfluenceRank algorithm (Figure 11.) is:

$$\boldsymbol{IR(i) = \left(1 - \delta(v)\right) * \sum_{j \in B_i} \frac{IR(j)}{N_j} + \delta(v) * MOI(v)}$$

where $\boldsymbol{N_j}$ is the total number of followers of node i, $\boldsymbol{IR(j)}$ is the InfluenceRank value of node *j*, $\boldsymbol{\delta(v)}$ is the popularity factor, and $\boldsymbol{MOI\ (i)}$ is the Magnitude of Influence of node *i*. This study evaluates the InfluenceRank algorithm using a Twitter dataset and contends that their algorithm provides a more accurate way to identify influencer when comparing with the original PageRank.

PR8: Some researches consider combining text data with activity data to process a more comprehensive influencer identification approach. These papers argue that when identifying influencer analyzing text content to receive the sentiment, interest topic and so on is not enough. User activity such as the number of followers, the frequency of posts, and the tenure of user should also be considerate influencer identification. Chen et al. (2012) analyze text data to

identify the similarity of interests between users and also activity data to measure user influence based on their social media activities. In their study, they fist define a network G = (*V*, *E*), where *G* is a directed network, *V* represents the users as nodes, *E* represents the edges between followers and followees. The authors then define a Relative Influence ($\boldsymbol{RI}$) model as follows:

$$\boldsymbol{RI}(\boldsymbol{v_i}, \boldsymbol{v_j}) = \boldsymbol{Q}(\boldsymbol{v_i}) + \boldsymbol{R}(\boldsymbol{v_i}, \boldsymbol{v_j}) + \boldsymbol{Sim}(\boldsymbol{v_i}, \boldsymbol{v_j})$$

where $\boldsymbol{Q}(\boldsymbol{v_i})$ measures the content quality, $\boldsymbol{R}(\boldsymbol{v_i}, \boldsymbol{v_j})$ represents the retweet behavior, and $\boldsymbol{Sim}\ (\boldsymbol{v_i}, \boldsymbol{v_j})$ is the similarity of interest between user *i* and user *j*. These factors are explained as follows:

Content quality ($\boldsymbol{Q}(\boldsymbol{v_i})$) is measured by the ratio of the number of retweets and comments user *i* received to the total amount of posts.

$$\boldsymbol{Q}(\boldsymbol{v_i}) = \frac{\boldsymbol{Retweeted}(\boldsymbol{v_i}) + \boldsymbol{Commented}(\boldsymbol{v_i})}{\boldsymbol{Tweets}(\boldsymbol{v_i})}$$

$R(\boldsymbol{v_i}, \boldsymbol{v_j})$ is the ratio of the number of posts user *j* retweets from *i* to the total number of retweets user *i* received.

$$\boldsymbol{R}(\boldsymbol{v_i}, \boldsymbol{v_j}) = \frac{\boldsymbol{Retweeted}(\boldsymbol{v_i}, \boldsymbol{v_j})}{\boldsymbol{Retweeted}(\boldsymbol{v_i})}$$

Similarity of interest includes two functions: user Interest Tag function and Content Keyword function. Each user has an interest tags set $\boldsymbol{T}(\boldsymbol{v_i}) = \{\boldsymbol{t_{i1}}, \boldsymbol{t_{i2,}}\boldsymbol{t_{i3}} \dots \boldsymbol{t_{ik}} \dots \boldsymbol{t_{im}}\}$ and a content keywords set $\boldsymbol{W}(\boldsymbol{v_i}) = \{(\boldsymbol{k_{i1}}, \boldsymbol{w_{i1}}), (\boldsymbol{k_{i2}}, \boldsymbol{w_{i2}}), (\boldsymbol{k_{i3}}, \boldsymbol{w_{i3}}) \dots (\boldsymbol{k_{ik}}, \boldsymbol{w_{ik}}) \dots (\boldsymbol{k_{im}}, \boldsymbol{w_{im}})\}$. In the content keywords set $\boldsymbol{W}(\boldsymbol{v_i})$, $\boldsymbol{k_{ik}}$ is a keyword used by user $\boldsymbol{v_i}$ and $\boldsymbol{w_{ik}}$ is its sentiment weight. The Interest Tag function $\boldsymbol{TS}(\boldsymbol{v_i}, \boldsymbol{v_j})$ calculates the similarity of interest tags used by user *i* and *j*, and the Content Keywords function $\boldsymbol{KS}(\boldsymbol{v_i}, \boldsymbol{v_j})$ calculates the similarity of keyword weights. Based on the above two functions, the similarity of interest between user i and j is the combination of these two values.

$$\boldsymbol{TS}(\boldsymbol{v_i}, \boldsymbol{v_j}) = \frac{\sum_{\boldsymbol{k=1}}^{\boldsymbol{n}} \boldsymbol{t_{ik}} * \boldsymbol{t_{jk}}}{\sqrt{\sum_{\boldsymbol{k=1}}^{\boldsymbol{n}} \boldsymbol{t_{ik}^2} * \sum_{\boldsymbol{k=1}}^{\boldsymbol{n}} \boldsymbol{t_{jk}^2}}}$$

$$\boldsymbol{KS}(\boldsymbol{v_i}, \boldsymbol{v_j}) = \frac{\sum_{k=1}^{n} \boldsymbol{w_{ik}} * \boldsymbol{w_{jk}}}{\sqrt{\sum_{k=1}^{n} \boldsymbol{w_{ik}^2} * \sum_{k=1}^{n} \boldsymbol{w_{jk}^2}}}$$

$$\boldsymbol{Sim}(\boldsymbol{v_i}, \boldsymbol{v_j}) = \boldsymbol{TS}(\boldsymbol{v_i}, \boldsymbol{v_j}) + \boldsymbol{KS}(\boldsymbol{v_i}, \boldsymbol{v_j})$$

Next, the authors design the User Network Global Influence Rank (UNGI_Rank) modified from PageRank algorithm to identify the influencers from a dataset of microblog, Tencent Weibo. The UNGI_Rank is as follows:

$$\textbf{UNGI_Rank}\ (\boldsymbol{i}) = (\boldsymbol{1} - \boldsymbol{d}) \frac{\textbf{Followers}\ (\boldsymbol{i})}{\boldsymbol{N}} + \boldsymbol{d} \sum_{\boldsymbol{j} \in \boldsymbol{B_i}} \frac{\textbf{UNGI_Rank}\ (\boldsymbol{j}) * \boldsymbol{RI}(\boldsymbol{v_i}, \boldsymbol{v_j})}{\textbf{Followees}\ (\boldsymbol{j})}$$

where $\textbf{UNGI_Rank}(\boldsymbol{j})$ is the User Network Global Influence Rank value of node *j*, $\boldsymbol{N}$ is the total number of users, and $\boldsymbol{RI}(\boldsymbol{v_i}, \boldsymbol{v_j})$ is the Relative Influence power between user *i* and *j*. After comparing with other algorithms, the authors conclude that their UNGI_Rank algorithm receives a similar list of influencers with a low computation complexity.

PR9: Ma & Liu (2014) design a SuperedgeRank algorithm, including text data activity data into a modified PageRank algorithms. SuperedgeRank algorithm is based on the idea of "Supernetwork", which was originally proposed by Sheffi in 1985 (Sheffi, 1985). Supernetwork is defined as "networks that exists above and beyond existing networks (Nagurney & Dong, 2002; Nagurney & Wakolbinger, 2005) and are multi-layered, multi-leveled, multi-dimensional, multi-attributed, and have varying degrees of congestion and coordination. " Building upon the idea of supernetwork, Ma & Liu (2014) identify multiple layers in social media including social subnetwork, environment subnetwork, psychological subnetwork, and viewpoint subnetwork. These layers constitute a supernetwork linked by superedges ($\boldsymbol{SE}$). In their study, a social subnetwork refers to "the reply relation among users." An environment subnetwork refers to "the process of information dissemination." A psychological subnetwork refers to "the psychological classifications of users, which can be derived from their posts." A viewpoint subnetwork refers to "the keywords in the users' post." (Ma & Liu, 2014, p. 1359)

First of all, the authors introduce the measurement of environment subnetwork. Each environment node $\boldsymbol{e_i}$ (e.g., a thread, a topic, or a subject) provide different degree of influential power $\boldsymbol{I_{e_i}}$.

They argue that when the influential power is high, the environment node is more likely to be linked by superedges. The influential degree of $\boldsymbol{I_{e_i}}$ is determined by two indexes: breadth of information dissemination $\boldsymbol{B_{e_i}}$and depth of information dissemination $\boldsymbol{D_{e_i}}$.

$$\boldsymbol{I_{e_i} = B_{e_i} * D_{e_i}}$$

Breadth of information dissemination $\boldsymbol{B_{e_i}}$ is measured by the ratio of the connected superedge $\boldsymbol{F_{e_i}}$ (comments) of information node $\boldsymbol{e_i}$ (thread) to the total superedges $\boldsymbol{N}$ (comments) in the whole network.

$$\boldsymbol{B_{e_i} = \frac{F_{e_i}}{N}}$$

Depth of information dissemination of $\boldsymbol{D_{e_i}}$ is measured by the total frequency of this piece of in information node $\boldsymbol{e_i}$ in superedges (frequency of comments per user in this thread) and the number of users in social subnetwork affected by this information.

$$\boldsymbol{D_{e_i} = \frac{F_{e_i}/A_{e_i}}{N/N_a}}$$

where $\boldsymbol{A_{e_i}}$ is the number of users discussing in the information node $\boldsymbol{e_i}$, and $\boldsymbol{N_a}$ is the number of users in the whole network.

Secondly, psychological subnetwork is measured by the psychological tendency and psychological strength of posts. Psychological tendency is determined by the positive and negation direction of sentiment $\boldsymbol{p_i}$ of posts, and psychological strength is determined by the absolute value of sentiment $\boldsymbol{p_i}$. Hence, a psychological subnetwork factor is an integer in +1, -1. The idea that a high psychological correlation between two superedges means the high probability that information will transform in between. Following formula is the measurement of the psychological correlation between superedge $\boldsymbol{SE_i}$ and $\boldsymbol{SE_j}$.

$$\boldsymbol{p_{ij}} = \begin{cases} \boldsymbol{sign(p_i * p_j)/|p_i - p_j|}, & \boldsymbol{p_i \neq p_j} \\ \boldsymbol{1}, & p_i = \boldsymbol{p_j} \end{cases}$$

The viewpoint subnetwork is measure by the content similarity between two superedges $\boldsymbol{SE_i}$ and

$\boldsymbol{SE_j}$ based on the keywords used in their posts. With a high similarity between two superedges, there is more chance that two users will have mutual recognition and influence. The similarity between superedges $\boldsymbol{SE_i}$ and $\boldsymbol{SE_j}$ formula is as follows:

$$\boldsymbol{Sim(SE_i, SE_j) = Sim_{i,} = cos\,\theta = \frac{\sum_{k=1}^{m} w_{ik} * w_{jk}}{\sqrt{(\sum_{k=1}^{m} w_{ik}^2)(\sum_{k=1}^{m} w_{jk}^2)}}}$$

On the other hand, different from traditional PageRank using simple linkage, SuperedgeRank algorithm adapts the concept of Superedge Degree ($\boldsymbol{L_{se}}$) to replace the original out-degree measure. Superedge Degree ($\boldsymbol{L_{se}}$) is defined as "the number of other superedges with which a certain superedge is linked through its nodes." (Ma & Liu, 2014; J.-W. Wang et al., 2010) Based on above measurements, the users design a SuperedgeRank algorithms modified from PageRank algorithm as follows:

$$\boldsymbol{SuperedgeRank(SE_i) = \frac{1 - I_{e_i}}{N} + I_{e_i} \sum_{SE_j} \frac{SuperedgeRank(SE_j) * p_{ij} * Sim_{ij}}{L_{SE_j}}}$$

where *N* is the total superedge in the network, $\boldsymbol{I_{e_i}}$ is the degree of influential power of environment node $\boldsymbol{e_i}$, $\boldsymbol{SuperedgeRank(SE_j)}$ is the SuperedgeRank value of superedge $\boldsymbol{SE_j}$, $\boldsymbol{p_i}$ is the psychological correlation between superedge $\boldsymbol{SE_i}$ and $\boldsymbol{SE_j}$, and $\boldsymbol{Sim_{ij}}$ is content similarity between two superedges.

This study uses an isolate strategy to evaluate the SuperedgeRank algorithm. They first identify the influencers from the whole network and then take out these nodes from the network. After comparing the difference between with and without the influencers, the authors contend the SuperedgeRank algorithm successfully identifies the influencers from a Chinese online forum.

3. Centrality-based Mechanisms

CE1: Wei & Hong (2013) investigate a Chinese microblog site (Sina) using Degree Centrality to identify influencer. In this study, they use following behavior as the network ties between individual users to define a network G=(V, E), where V is the set of users, and E is the set of ties between followers and followees (edges). For example, if user *i*

follows user *j*'s content, there is one directional tie from *j* to *i* representing information flows from *j* to *i*. Based on this network G, the authors measure Degree Centrality of each users to identify influencers as the most dominant nodes in their social network. At the same time, this study also measures user activities such as number of articles, following nodes, and followers to analyze the correlation between activities of influencers and their followers and supports their argument that influencers' behavior will positively influence their followers.

Degree Centrality is the simplest measure of Centrality (Borgatti, Everett, & Johnson, 2013). Degree Centrality simply calculates the total number of ties connected to one node, including number of follower ties and followee ties (Freeman, 1977). The Degree Centrality of node ν is:

$$\boldsymbol{d_d(v)} = \sum_{\boldsymbol{i=1, i \neq v}} \boldsymbol{e(i, v)}$$

where $\boldsymbol{e(i, v)}$ is an edge directly connecting to node ν, and the normalized Degree Centrality is divided by the maximum possible degree (N-1) (Everett & Borgatti, 1999). The normalized Degree Centrality of node ν is a value ranging from 0 to1 as following:

$$\boldsymbol{d_{\bar{d}}(v)} = {\textstyle \sum_{\boldsymbol{i=1, i \neq v}} \boldsymbol{e(i, v)}} / \boldsymbol{N - 1}$$

CE2: Beside Degree Centrality, there are two other common Centrality measures: Betweenness Centrality and Closeness Centrality. Betweenness Centrality is "a measure of how often a given node falls along the shortest path between two other nodes." (Borgatti, Everett, & Johnson, 2013, p. 174) The Betweenness Centrality of the node ν is:

$$\boldsymbol{d_b(v)} = \sum_{\boldsymbol{i<j}} \frac{\boldsymbol{g_{ij}(v)}}{\boldsymbol{g_{ij}}}$$

Where i, j are two nodes in the same network as node ν, and $g_{ij}(v)$ is the total number of shortest paths connecting *i* and *j* through ν, and g_{ij} is the total shortest paths connecting *i*

and j. The basic idea of Betweenness Centrality is to emphasize the broker feature by measuring the chance that a give node is required for two other nodes to reach each other by the shortest path. When the degree of Betweenness Centrality is high, the nodes represent the more critical positions in their network to bridge information flows (D. R. White & Borgatti, 1994). The normalized Betweenness Centrality of node ν is divided by the number of pairs of nodes, which is not including node ν. The number of pairs of nodes is calculated by (N-1)(N-2) for a directed network and (N-1)(N-2)/2 for an undirected network, where N is the number of nodes in the network (D. R. White & Borgatti, 1994). The normalized Betweenness Centrality of the node ν is:

Directed Network

$$\boldsymbol{d_{\bar{b}}(v)} = \sum_{i<j} \frac{\boldsymbol{g_{ij}(v)}}{\boldsymbol{g_{ij}}} \Big/ \boldsymbol{(N-1)(N-2)}$$

Undirected Network

$$\boldsymbol{d_{\bar{b}}(v)} = \sum_{i<j} \frac{\boldsymbol{g_{ij}(v)}}{\boldsymbol{g_{ij}}} \Big/ \boldsymbol{(N-1)(N-2)/2}$$

Closeness Centrality is the sum of the lengths of the shortest paths from one node to all other nodes. Closeness is an inverse measurement that the node with smaller value indicates a more central position in the network. A more central node can diffuse information to all other nodes more easily because of the shorter traveling distance to all other nodes (Borgatti, 2005; Freeman, 1979). The Closeness Centrality of node ν is given by

$$\boldsymbol{d_c(v)} = \sum_{\boldsymbol{i=1 \neq v}} \boldsymbol{l(i,v)}$$

where $l\ (i, v)$ is the shortest path from i to ν, and the normalized Closeness Centrality is by the maximum possible degree (N-1), where N is the total number of nodes (Borgatti,

Everett, & Johnson, 2013). The Closeness Centrality of node ν is a value ranging from 0 to1:

$$\boldsymbol{d_{\bar{c}}(v)} = {\textstyle\sum_{i=1\neq v} \boldsymbol{l(i,v)}} / \boldsymbol{N-1}$$

Liu, Yu, & Lu (2013) adopt all three Centrality measures we discussed earlier in identifying influencer. This research designs a Synthesis Centrality (SC) measurement to value each node. The Synthesis Centrality method combining multiple centrality values to rank each node v by the following formula

$$\boldsymbol{SC(v)} = \frac{\boldsymbol{d_{\bar{d}}(v)} + \boldsymbol{d_{\bar{b}}(v)}}{\boldsymbol{d_{\bar{c}}(v)}}$$

where $d_{\bar{d}}(v)$ is the normalized Degree Centrality, $d_{\bar{b}}(v)$ is the normalized Betweenness Centrality, and $d_{\bar{c}}(v)$ is the normalized Closeness Centrality.

Based on the Synthesis Centrality (SC) measurement, the authors identify the top 20 influencers and compare the results with original HITS algorithm and PageRank algorithm. Their experiment results supports that Synthesis Centrality (SC) provide a higher accuracy for influencer identification.

4. Clustering-based Algorithms

CL1: Hudli et al. (2012) define eight attributes from user activities such as the time user spends online, or the frequency one user posts content or replies to another user, and from the text features of post such as sentiment polarity, or the average length of contents. Based on these attributes, the authors employ K-means clustering algorithm to analyze a discussion forum dataset. From five different types of discussion forums (consumer product, travel, technology, healthcare, and entertainment), this study identifies 10 to 20 percent of users as influencers from each discussion forum. The authors argue that these influencers will be the niches for marketing strategy targeting.

CL2: some researches put the text content feature into consideration for influencer identification. They adopt sentiment analysis and text mining and include the results when pre-defining the attributes of influencer. Clustering algorithm is then employed to analyze these attributes for influencer identification. Duan at al. (2014) apply the clustering

algorithm with sentiment analysis to identify opinion leader s from a web-based stock message forum. Their study first define attribute based on user activities (e.g., number of posts and replies) to employ a fuzzy-based method in the K-group clustering algorithm. The authors then use the top K groups of users as the influencer candidates. Next, they analyze the sentiment polarity of posts of influencer candidates and compare with the actual stock price movement. Finally, the influencers are identifies by the correlation between their post sentiment and actually stock price movement. Based on the comparison of average correlation coefficient, the authors contend their method is more accurate than the Pagerank-based method.

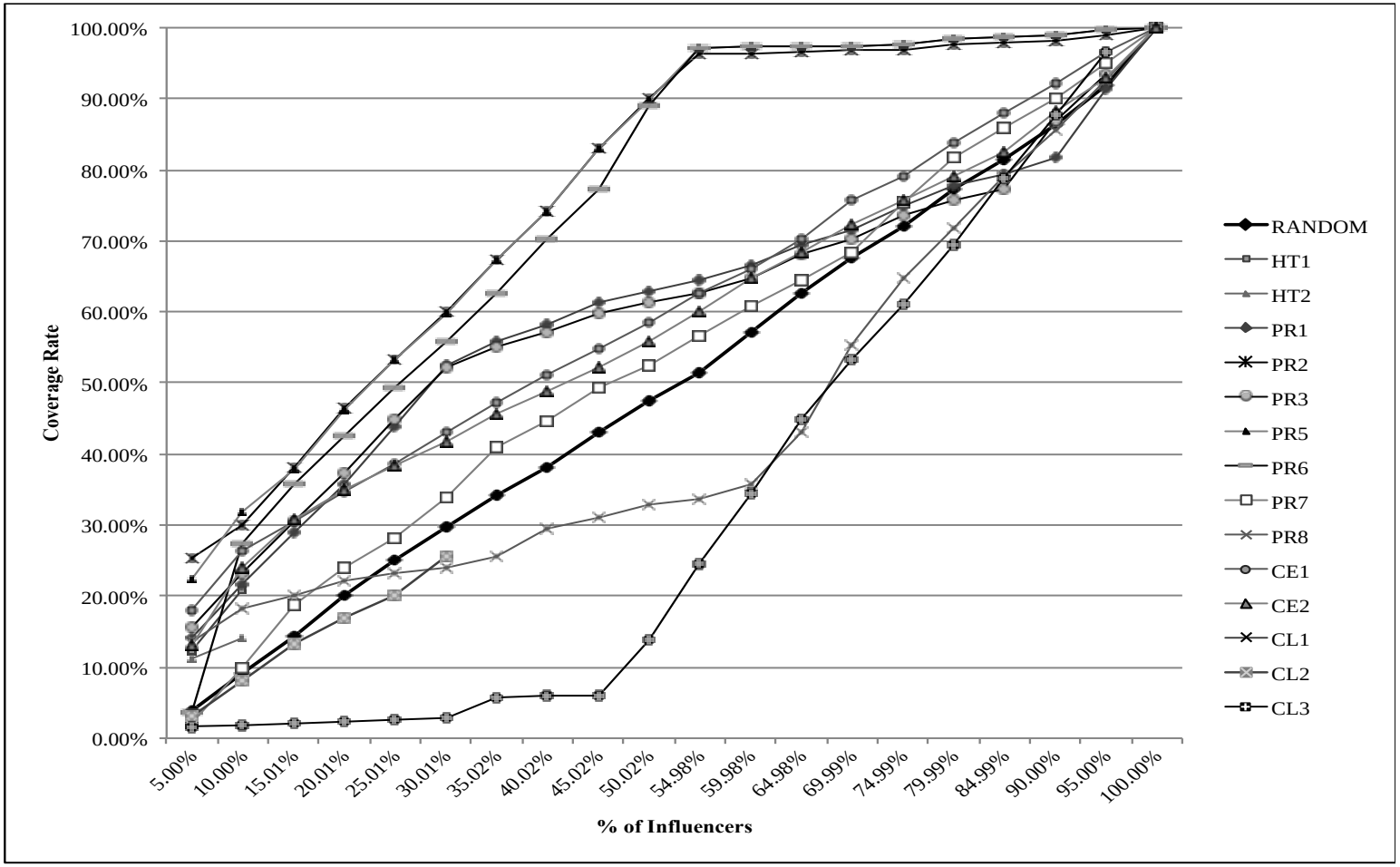
CL3: Some studies further include social network attributes to analysis the social media comment network, which the linkage between users is weaved by comment or reply. Incorporated with text feature and network attributes, researchers improve the clustering algorithm in influencer identification. Song et al. (2011) define the comment network based on explicit link and implicit link. Following or replying behavior is counted as explicit link, and sentiment similarity is counted as implicit link. Meanwhile, the explicit/ implicit links can be detailed as positive or negative link based on the sentiment orientation toward contents. The authors adapt a PageRank-liked algorithm called Dynamic OpinionRank algorithm measuring the text content to estimate the Comment Quality. Further, this study calculates the Degree Centrality and Proximity Prestige of each node and combines them with the comment quality score. The authors then use a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to cluster all the nodes, and to identify the influencers from those outlier nodes. Based on the experiment result Sina news forums, the authors define the outliers from the clusters as the influencers.

Appendix II: Evaluation Metrics Review from Literature

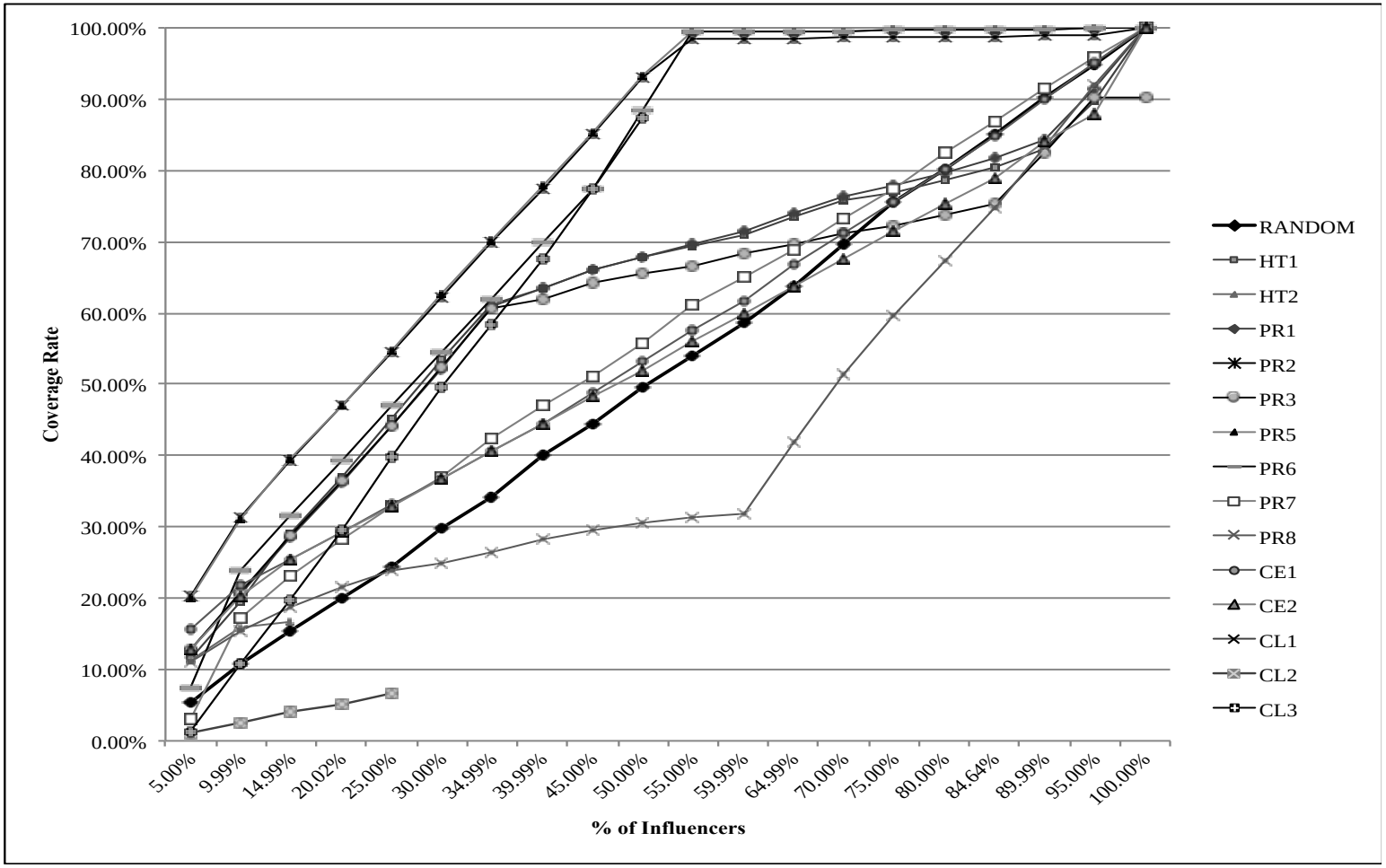
Code	Dataset Type	Data Source	# of Users	# of Posts	# of Threads	Data Collecting Approach	Quality Assessment
HT1	Microblog	Sina				Use keyword to select posts from 01/01/2014 to 02/28/2014	Compare influencers identified from algorithm with those from human beings
HT2	Microblog	Twitter	18,713	44,391		Use keywords to collect tweets related to the UK General Election in 2010	Precision, Recall, F-measure
PR1	Review Site	Epinions.com		49,471		Collect reviews from four different categories (i.e., Digital, Movie, Fax, and Travel)	Measure the similarity of score between the identified influencer and a trust rank list from the website.
PR2	Forum	Sina	206	1,481		Use keyword to crawl posts from 01/01/2011 to 12/31/2011	Artificially evaluate the identified influencers
PR3	Forum	CCNU BBS	2,215	19,687		Collect data from 120 popular boards between 01/01/2006 to 10/01/2009 and select the biggest one to analyze	Use core ratio to calculate the frequency of interaction between influencers and others.
PR4	Forum	Tianya Club	374,302	357,283		Collect from the Tianya BBS	Compare the identified influencers with the ranking provided from the website
PR5	Forum	Sina	106			Collect from the Sina web from 05/01/2012 to 12/31/2012	Calculate the coverage of top N percent influencers
PR6	Forum	SMTH forum	21,725	284,443	33,883	Select 34 popular boards from SMTH BBS between 04/01/2008 and 05/01/2008	1. Use coverage ratio calculating the followers of influencers 2. Compare the influencers based on the user category provided from the website

Code	Dataset Type	Data Source	# of Users	# of Posts	# of Threads	Data Collecting Approach	Quality Assessment
PR7	Microblog	Friend Feed 2010 dataset	665,000	80,000,000		Use the Friend Feed 2010 dataset	Use PageRank and Limited Recursive Algorithm (LRA) to evaluate the identified influencers
PR8	Microblog	Tencent Weibo		2,320,895		A public dataset from the 2012 KDD CUP (Data Challenge)	Compare the identified influencers with the results from the TunkRank algorithm
PR9	Forum	Tianya Club	671	1,019		Use keyword to collect posts from 03/17/2011 to 03/18/2011	Design a method based on the mean average precision approach to measure the importance of identified influencers
CE1	Microblog	Sina	120			Randomly select microblogger from 25 province in 06/27/2013	Correlate each influencer with number of followees, followers, and articles
CE2	Microblog	Sina	4,356			Select users based on the students form Shanghai university	Compare results with PageRank and HITS algorithms
CL1	Forum		5,850			Collect data from discussion forum including five different topics	
CL2	Forum					Collect four years data from a stock forum	Measure the correlation coefficient between identified influence with the results from the PageRank algorithm based on the stock prediction
CL3	Forum	Sina		118		Use keyword to collect posts from 03/17/2011 to 03/18/2011	F-measure

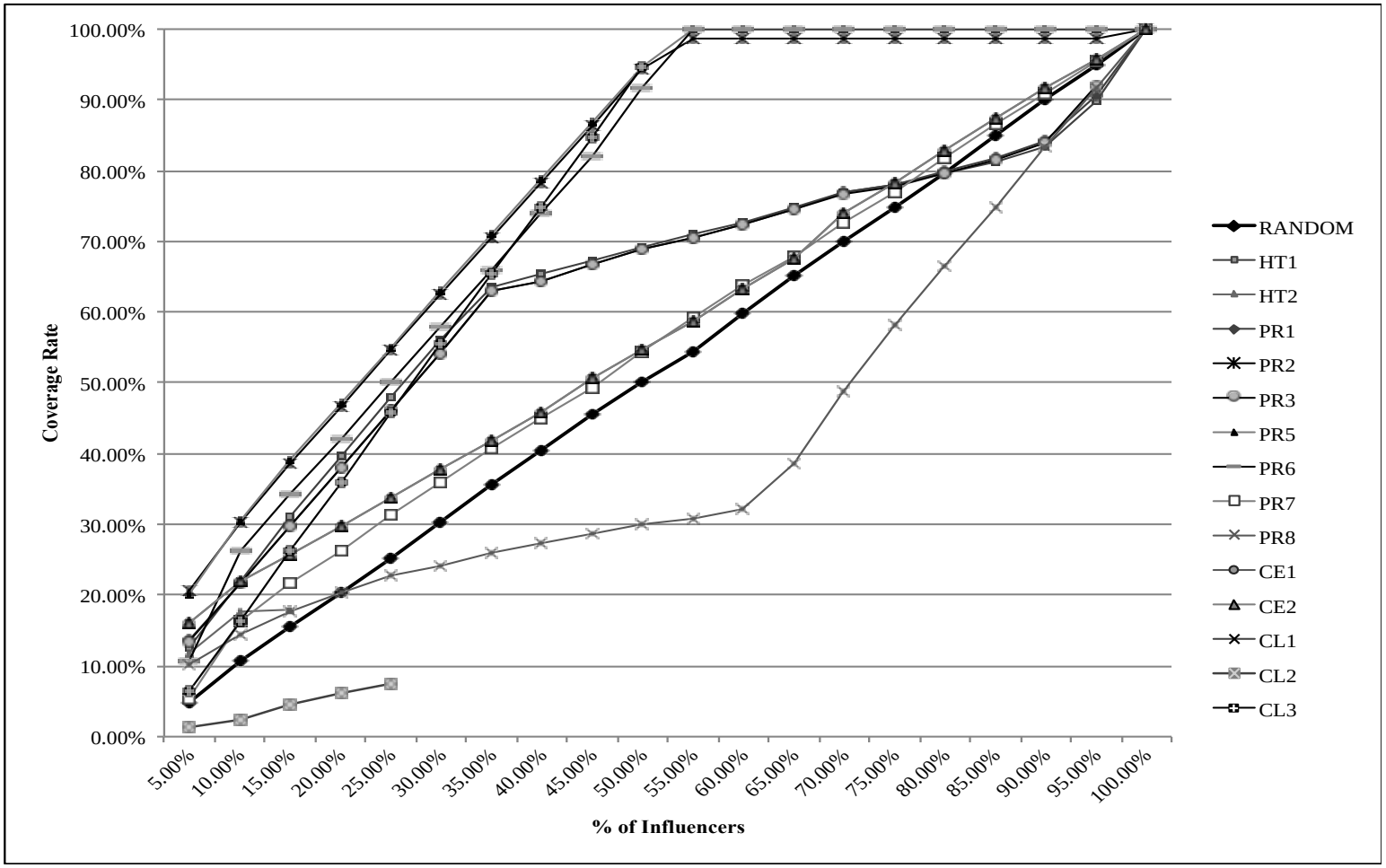
Appendix III: Experiment Results from the Twitter March Madness Datasets



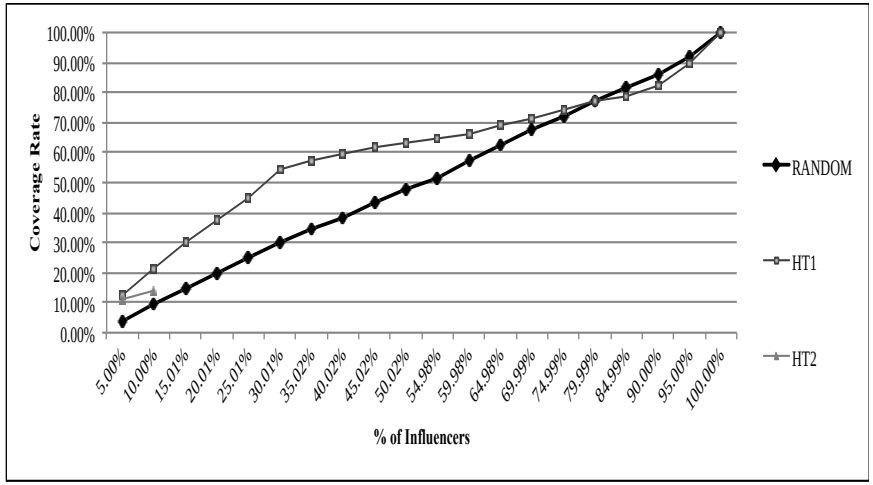
Coverage Rate for Different Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



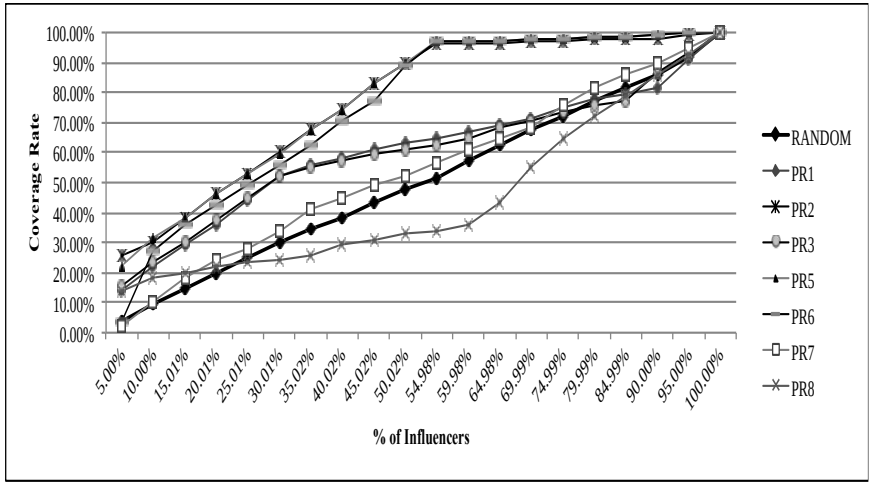
Coverage Rate for Different Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



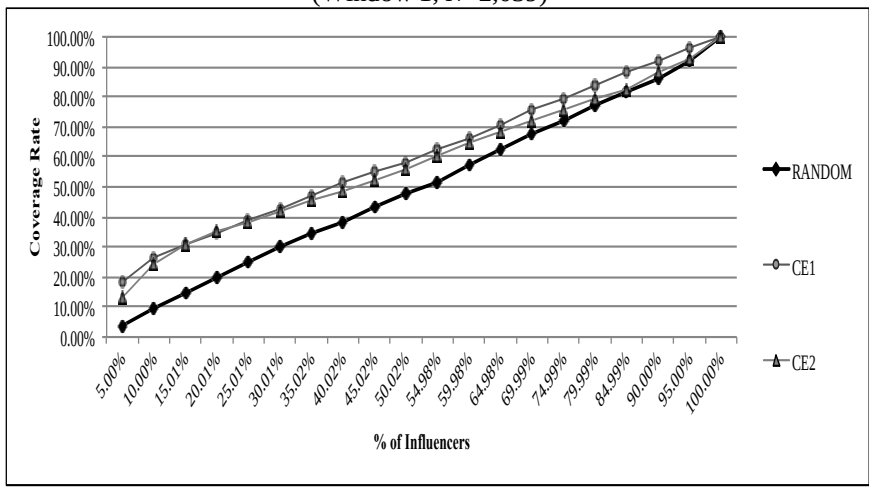
Coverage Rate for Different Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



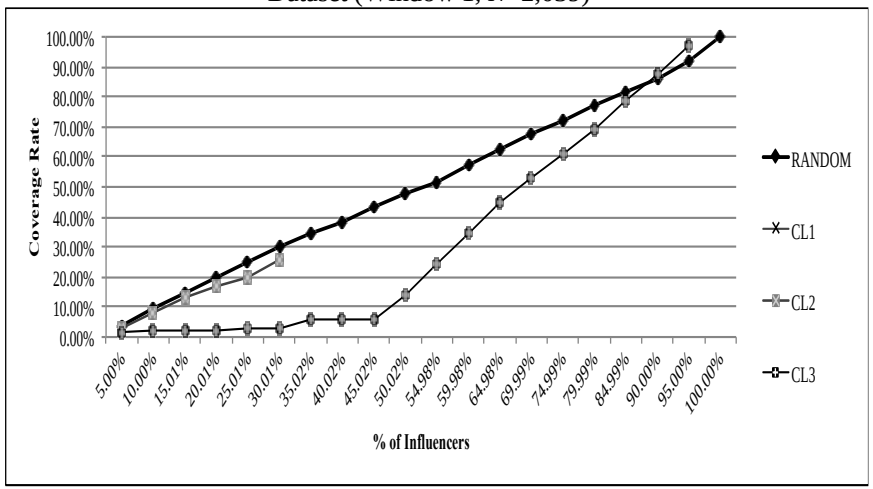
Coverage Rate for HITS-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



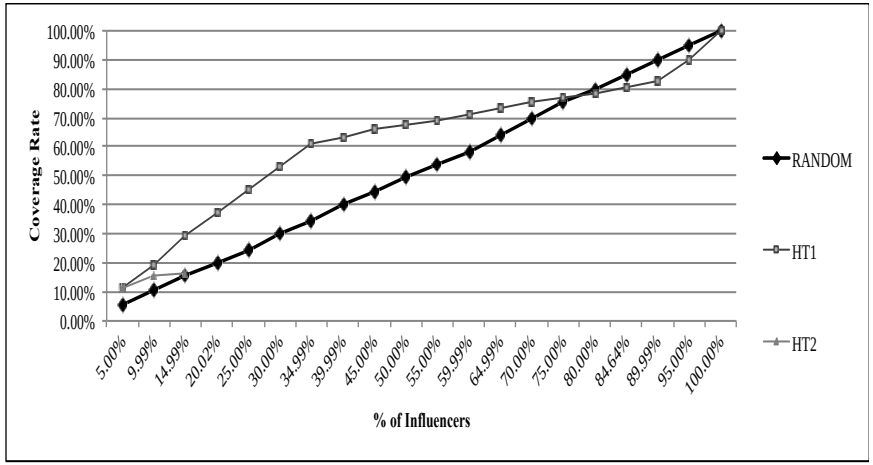
Coverage Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



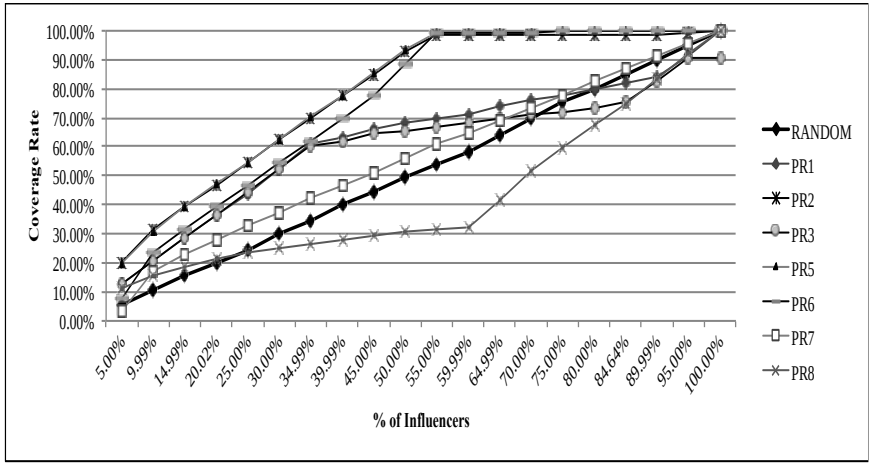
Coverage Rate for Centrality-based Methods: Twitter March Madness Dataset (Window 1, *N*=2,039)



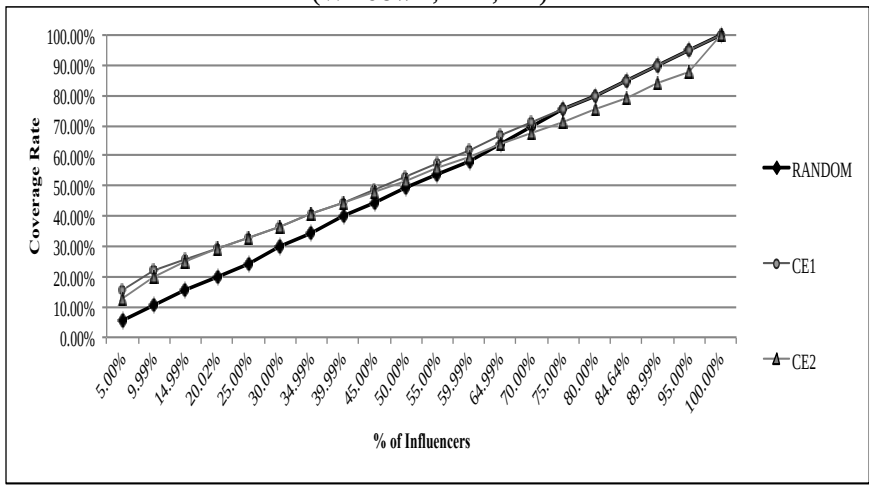
Coverage Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



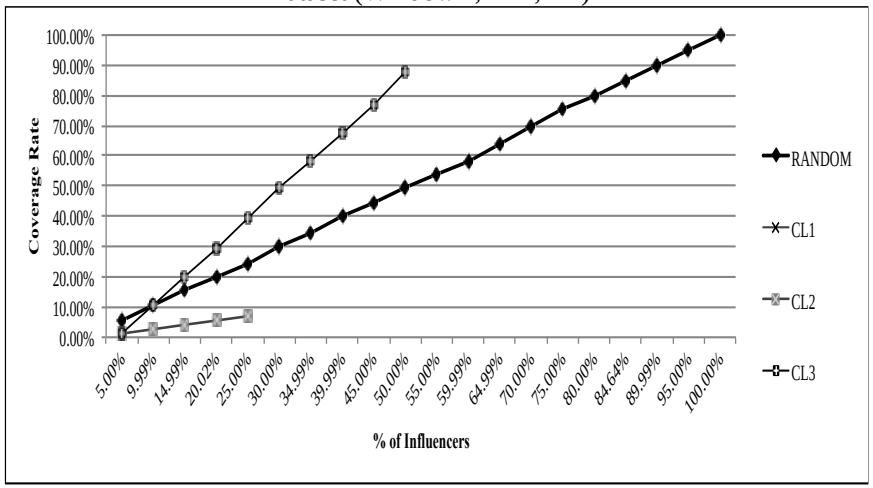
Coverage Rate for HITS-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



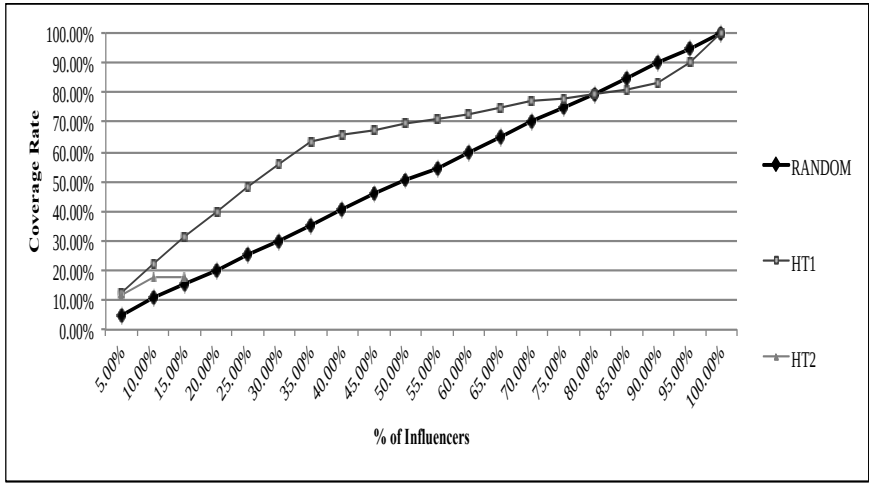
Coverage Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



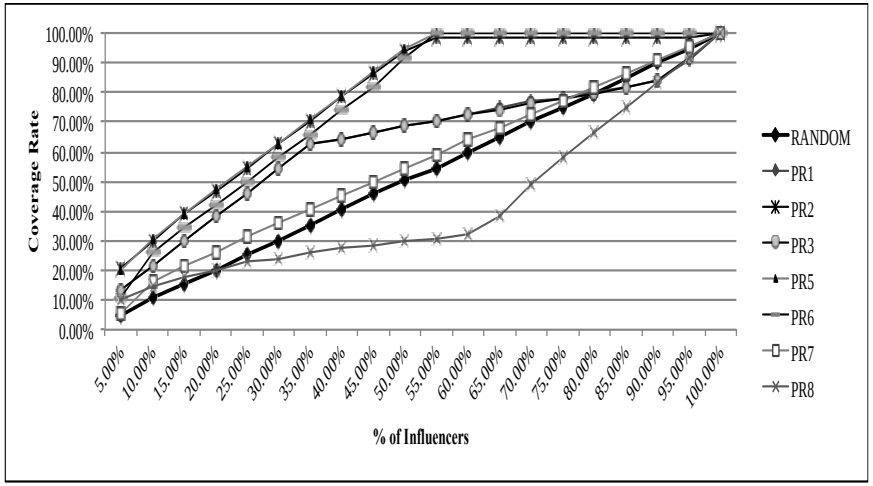
Coverage Rate for Centrality-based Methods: Twitter March Madness Dataset (Window 2, *N*=7,424)



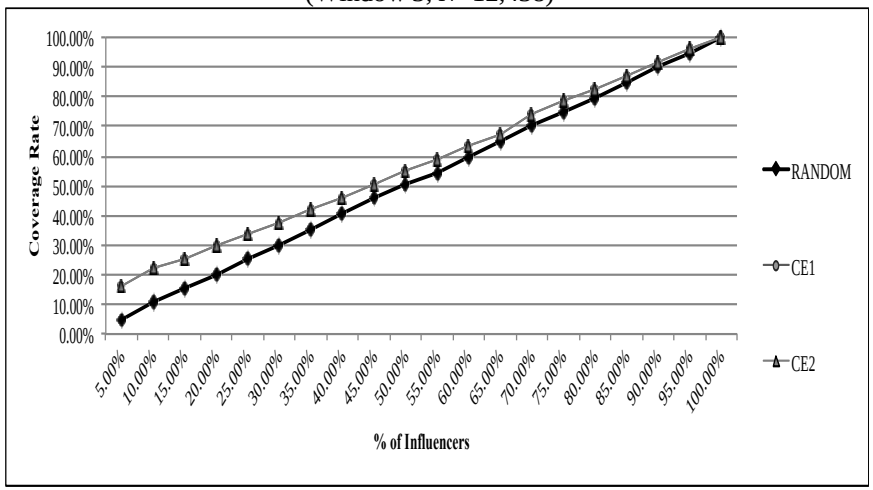
Coverage Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



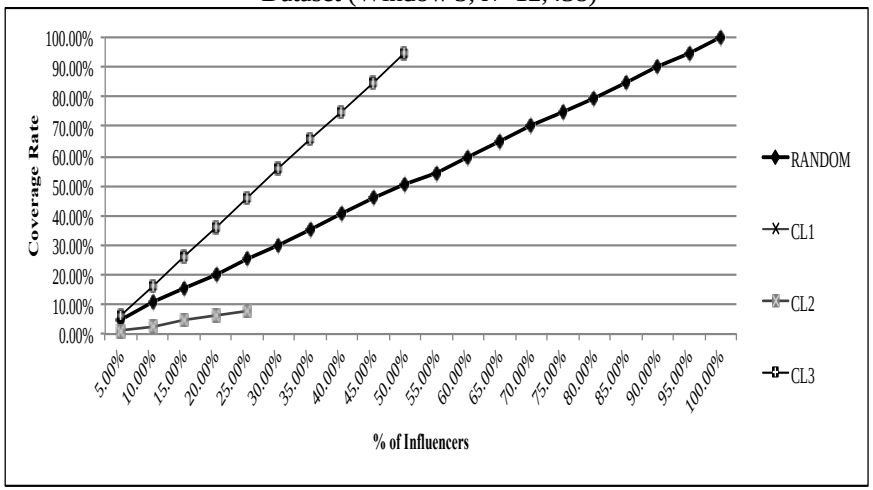
Coverage Rate for HITS-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



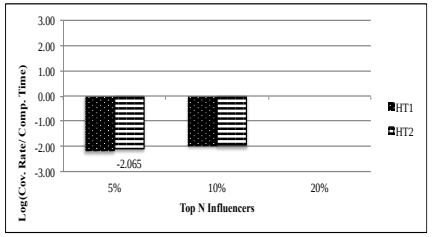
Coverage Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



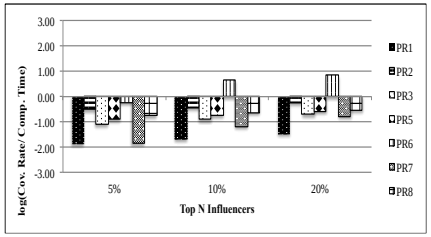
Coverage Rate for Centrality-based Methods: Twitter March Madness Dataset (Window 3, *N*=12,438)



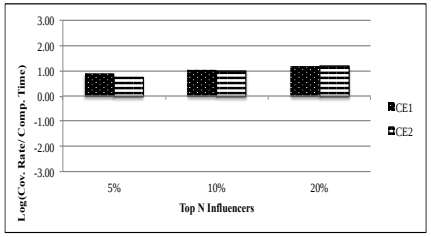
Coverage Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



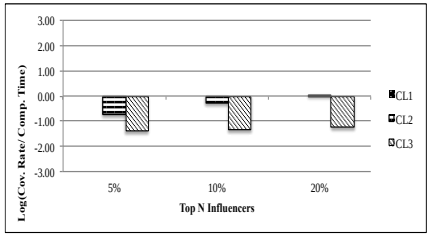
Coverage Rate/ Computation Time for HITS-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



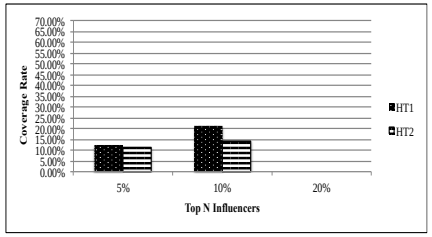
Coverage Rate/ Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



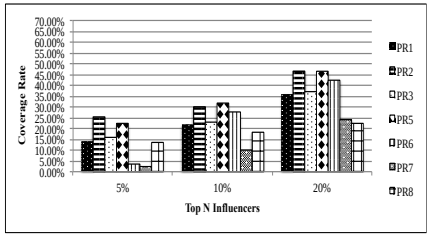
Coverage Rate/ Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 1, *N*=2,039)



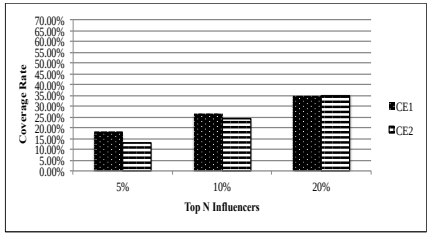
Coverage Rate/ Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



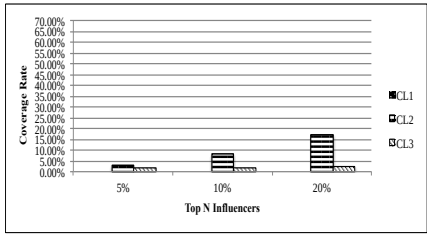
Coverage Rate for HITS-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



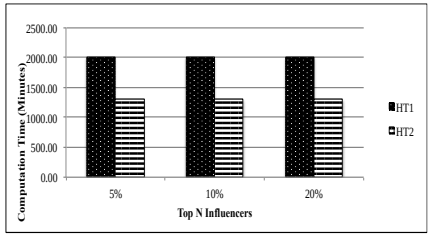
Coverage Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



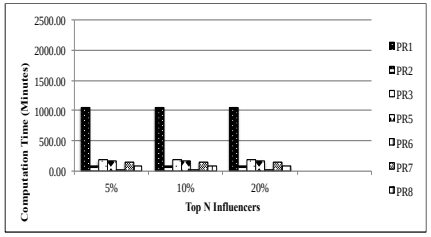
Coverage Rate for Centrality-based Methods: Twitter March Madness Dataset (Window 1, *N*=2,039)



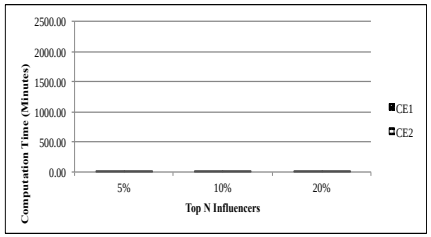
Coverage Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



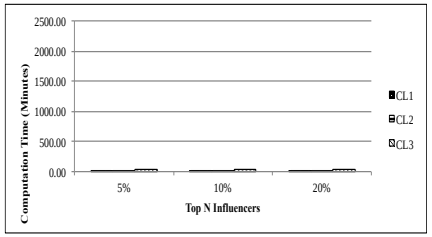
Computation Time for HITS-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



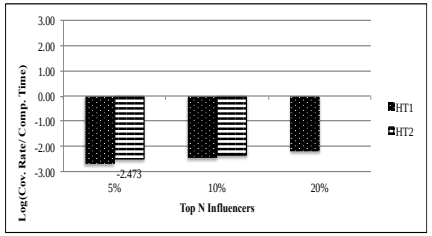
Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



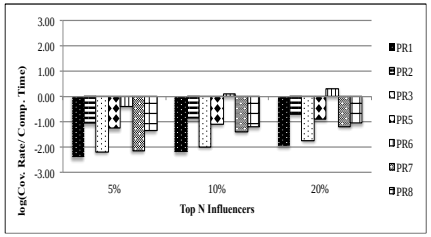
Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 1, *N*=2,039)



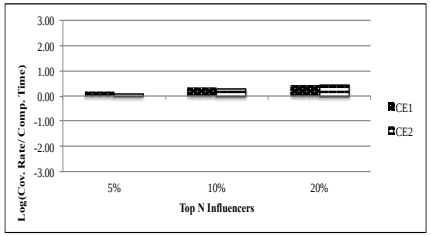
Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



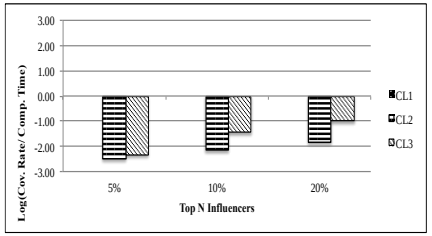
Coverage Rate/ Computation Time for HITS-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



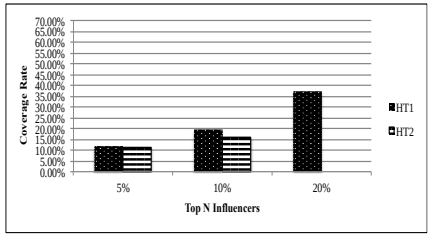
Coverage Rate/ Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



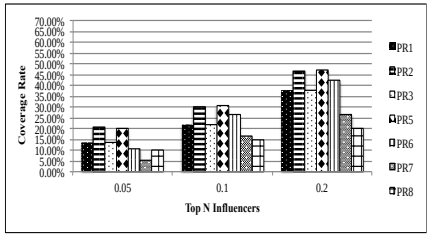
Coverage Rate/ Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 2, *N*=7,424)



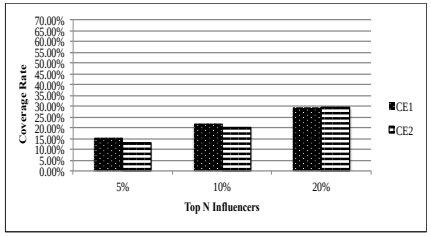
Coverage Rate/ Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



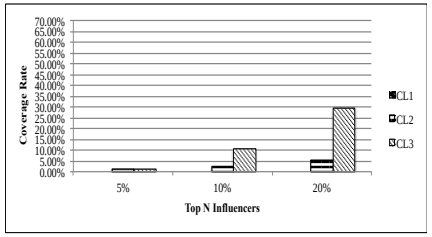
Coverage Rate for HITS -based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



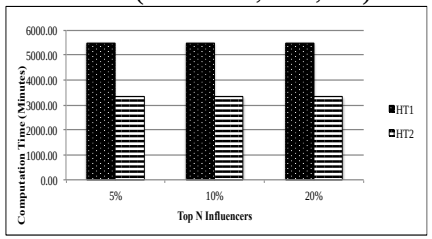
Coverage Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



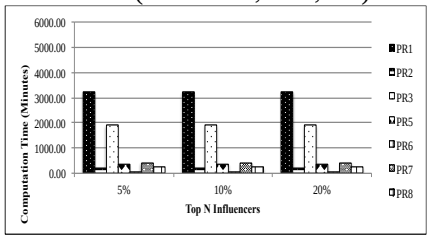
Coverage Rate for Centrality-based Methods: Twitter March Madness Dataset (Window 2, *N*=7,424)



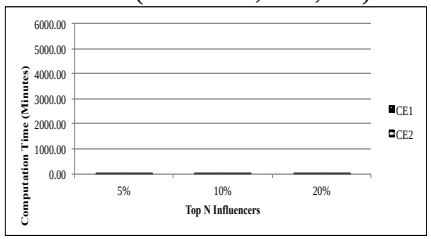
Coverage Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



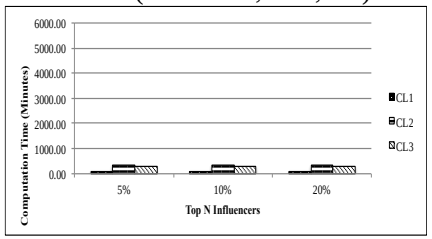
Computation Time for HITS -based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



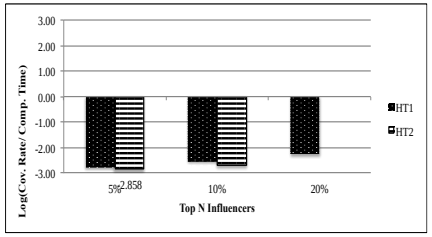
Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



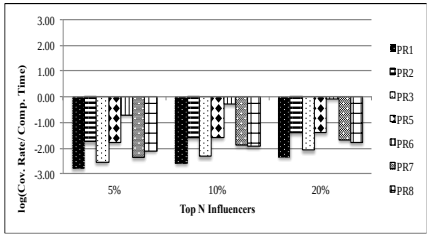
Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 2, *N*=7,424)



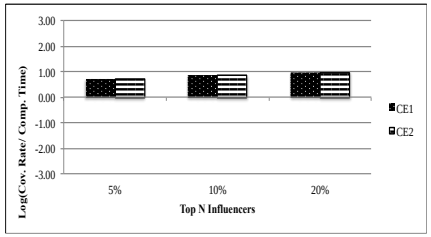
Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



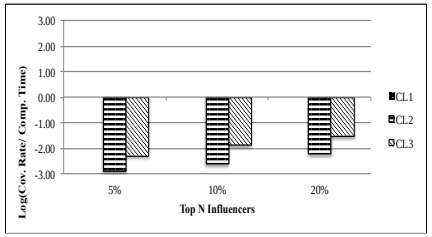
Coverage Rate/ Computation Time for HITS-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



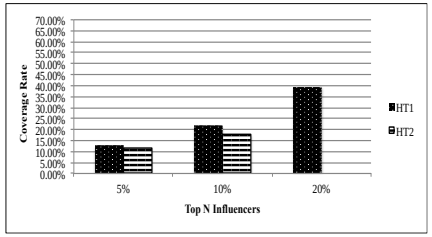
Coverage Rate/ Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



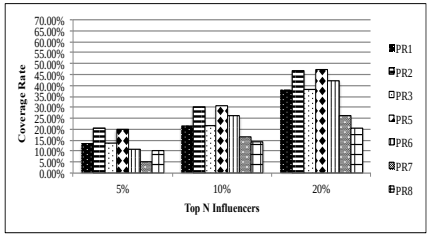
Coverage Rate/ Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 3, *N*=12,438)



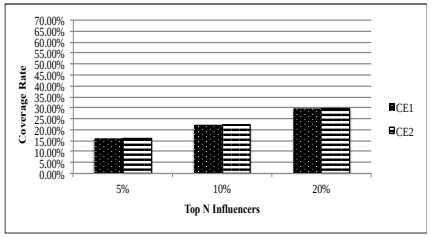
Coverage Rate/ Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



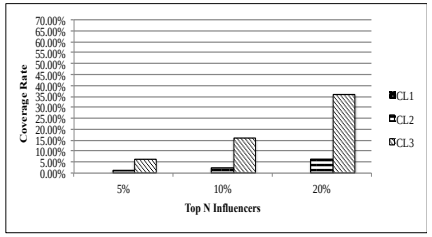
Coverage Rate for HITS -based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



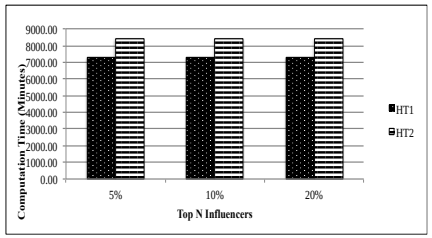
Coverage Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438))



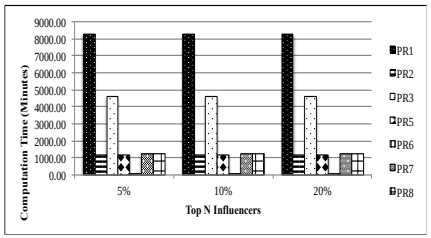
Coverage Rate for Centrality-based Methods: Twitter March Madness Dataset (Window 3, *N*=12,438)



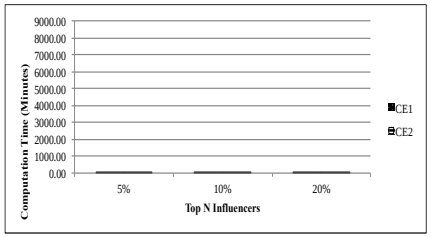
Coverage Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



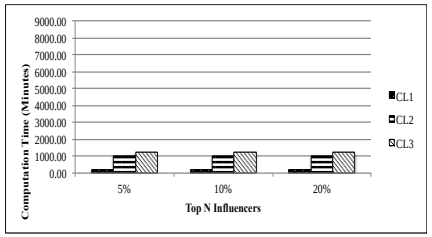
Computation Time for HITS -based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



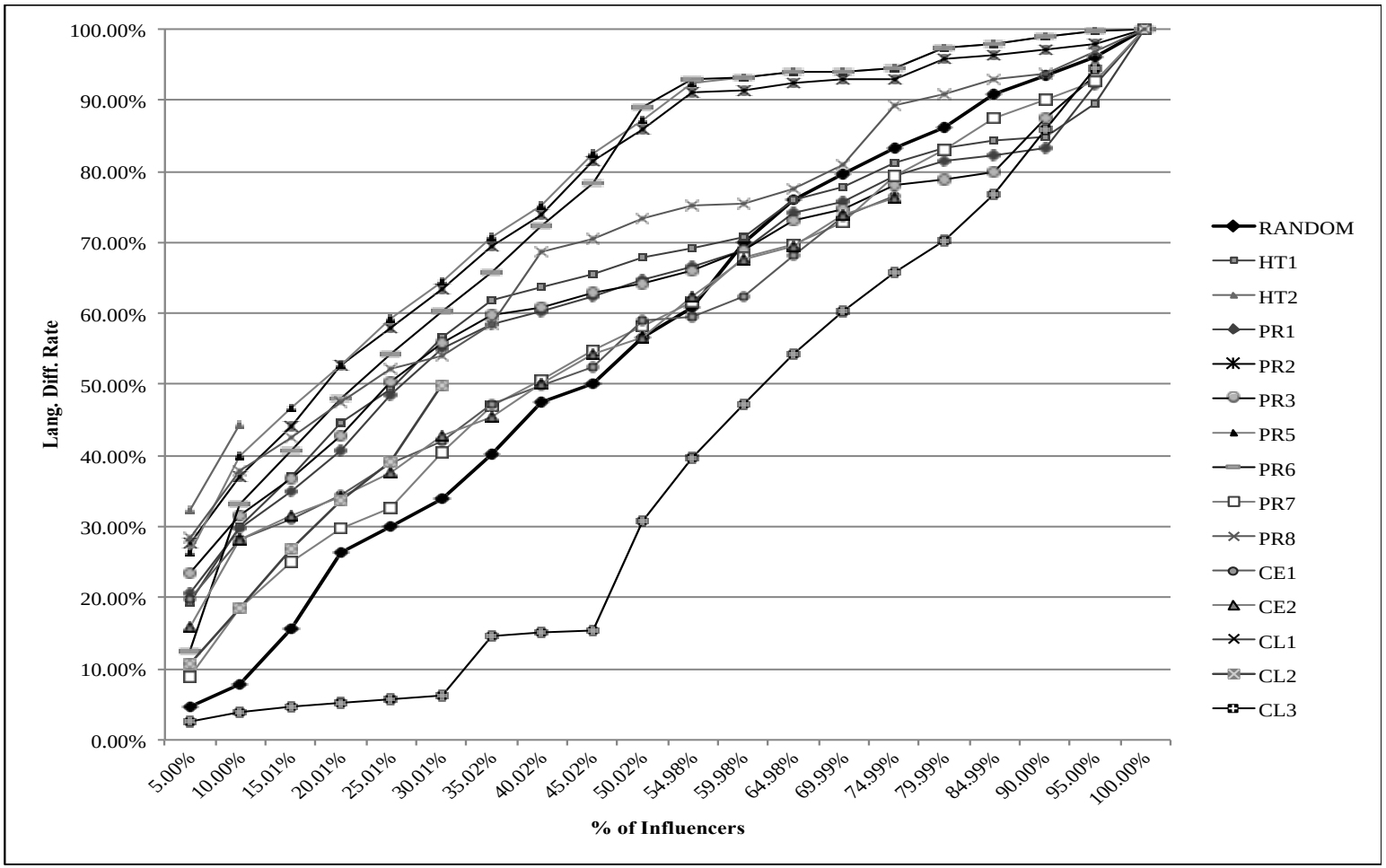
Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



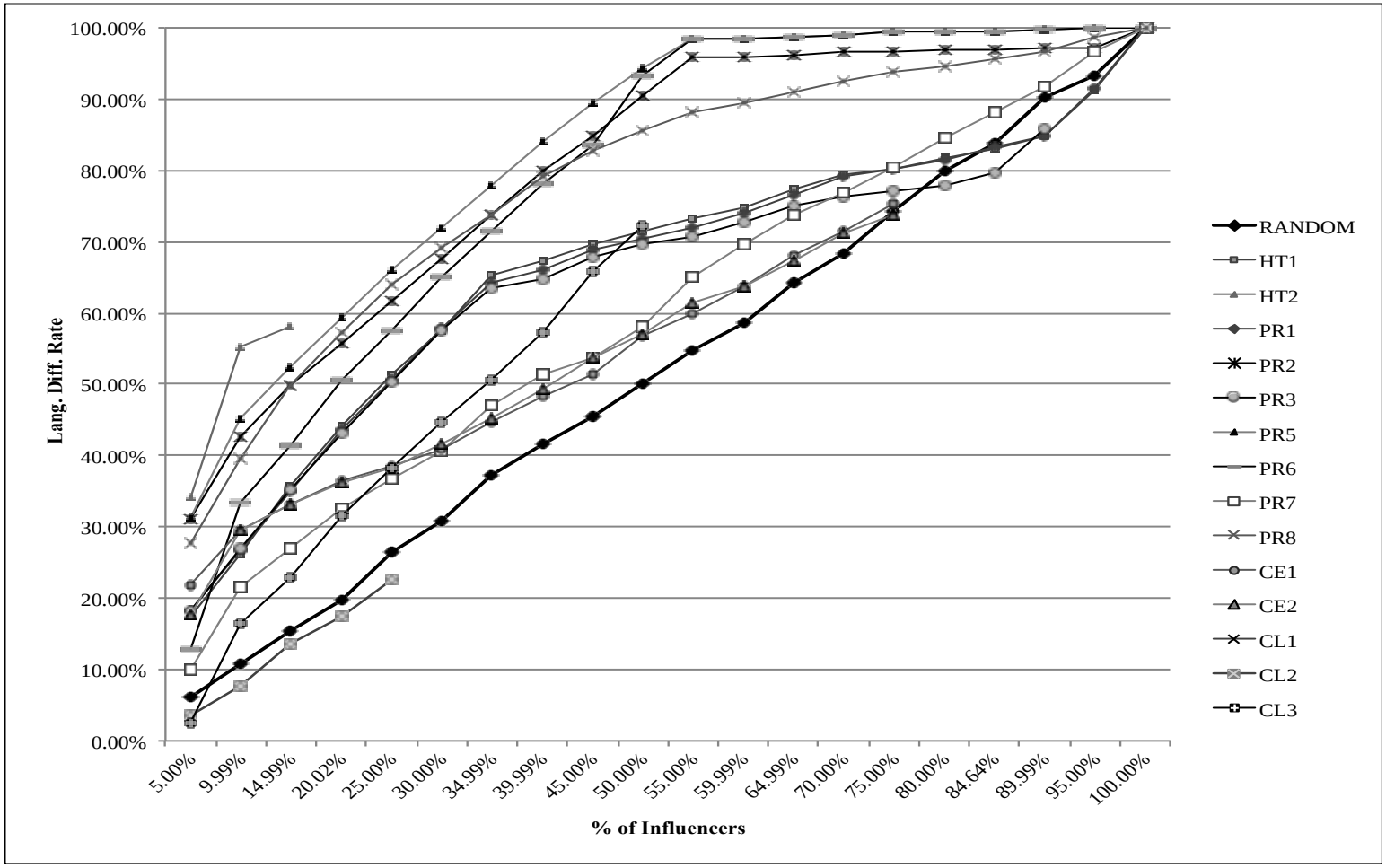
Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 3, *N*=12,438)



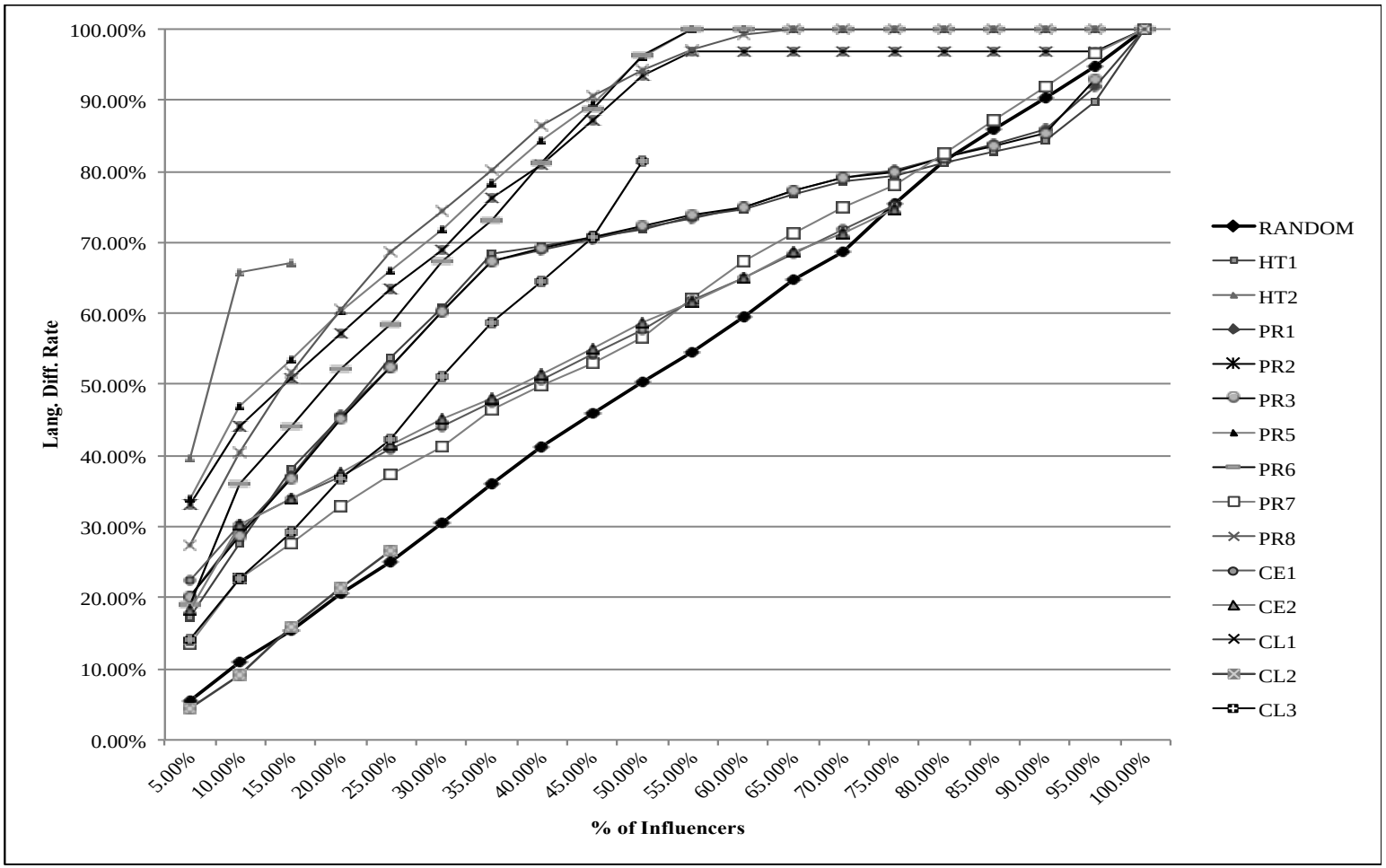
Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



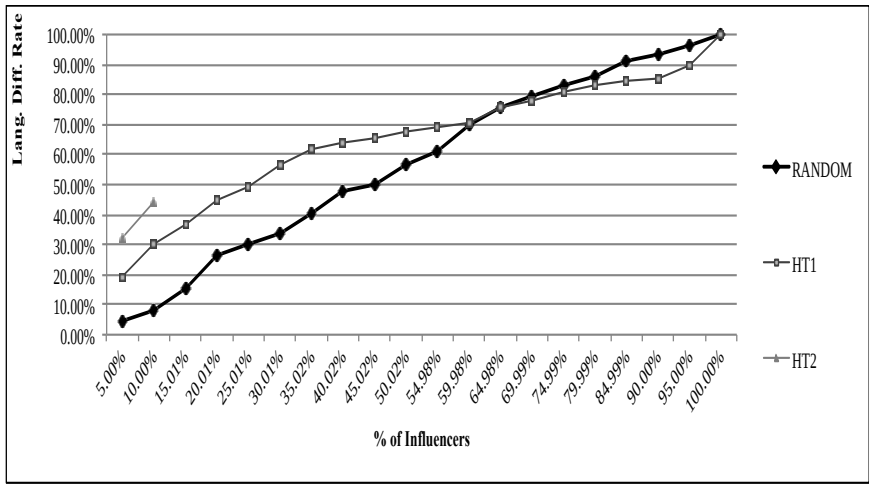
Language Diffusion Rate for Different Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



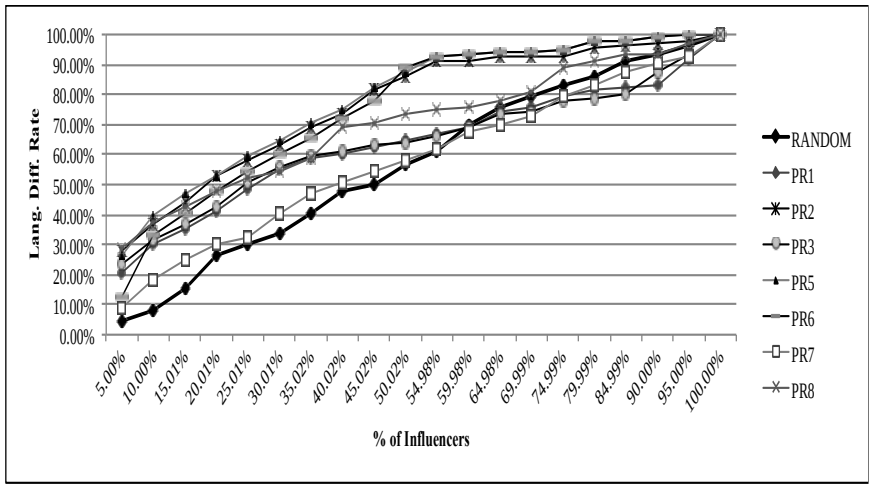
Language Diffusion Rate for Different Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



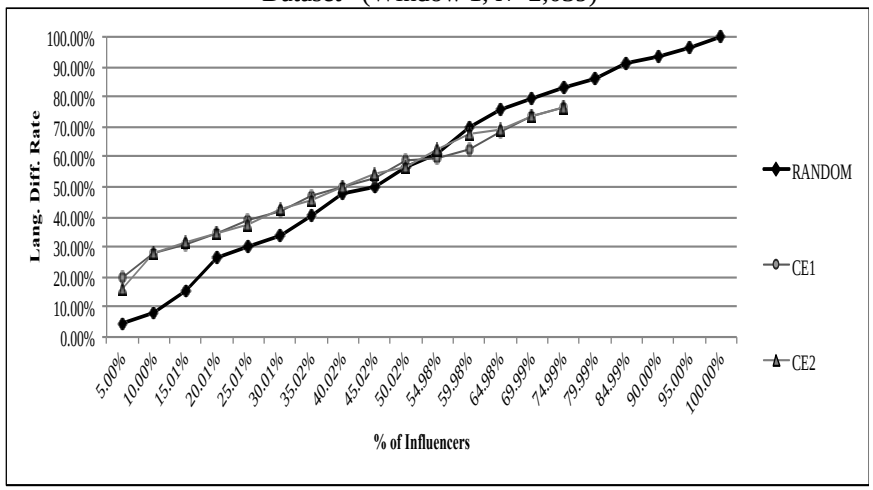
Language Diffusion Rate for Different Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



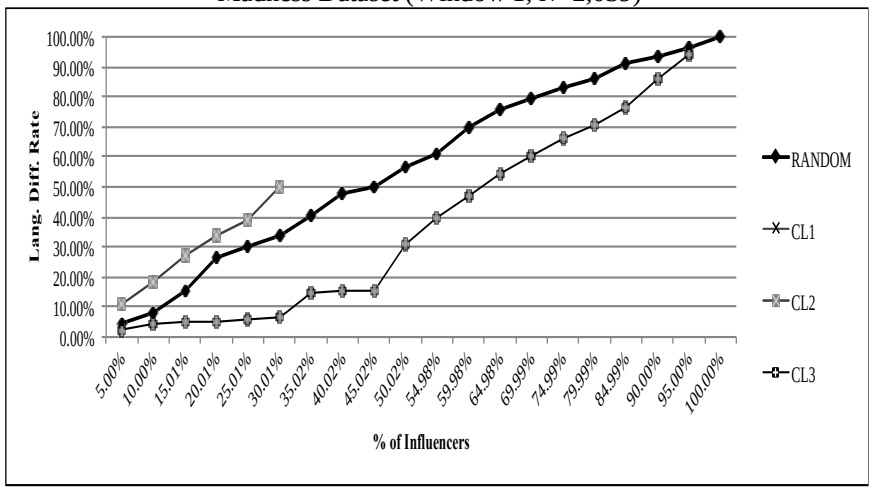
Language Diffusion Rate for HITS-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



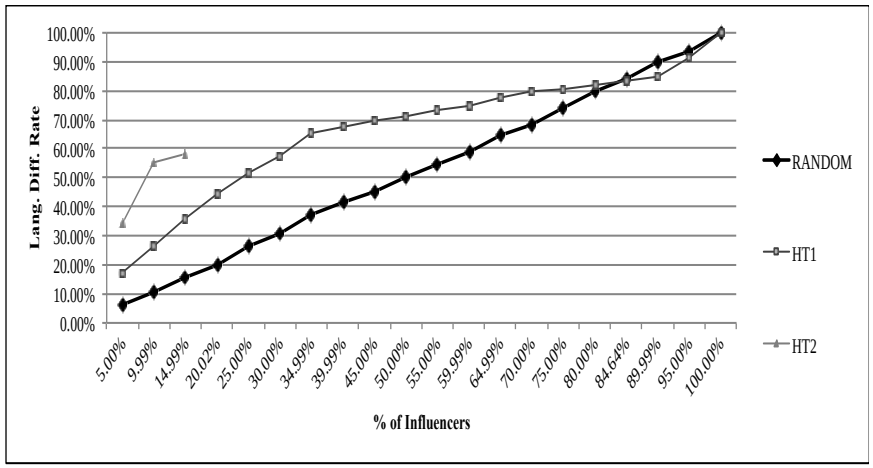
Language Diffusion Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



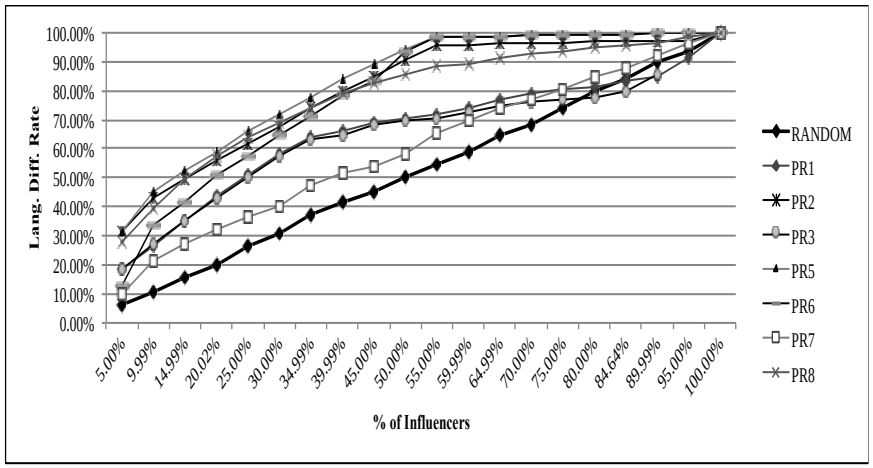
Language Diffusion Rate for Centrality-based Methods: Twitter March Madness Dataset (Window 1, *N*=2,039)



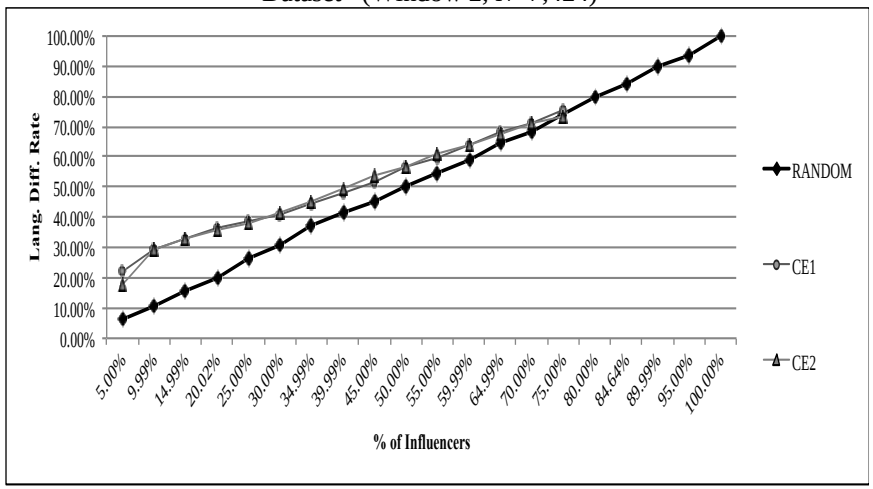
Language Diffusion Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



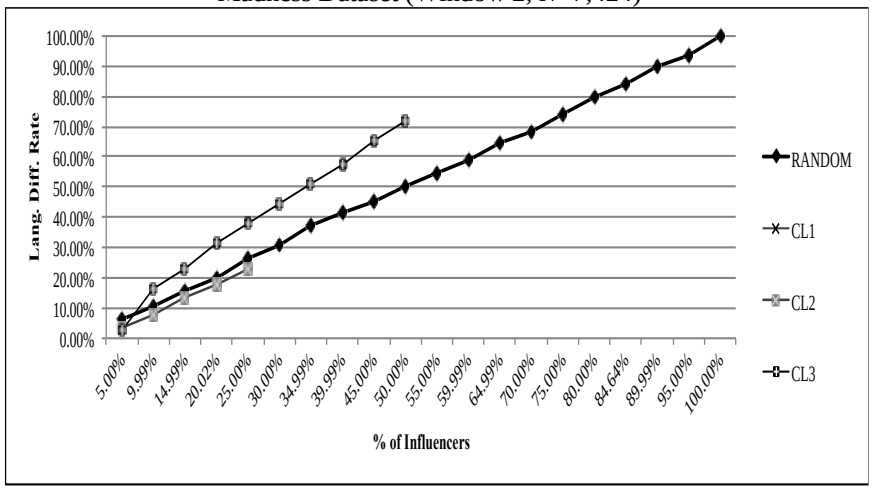
Language Diffusion Rate for HITS-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



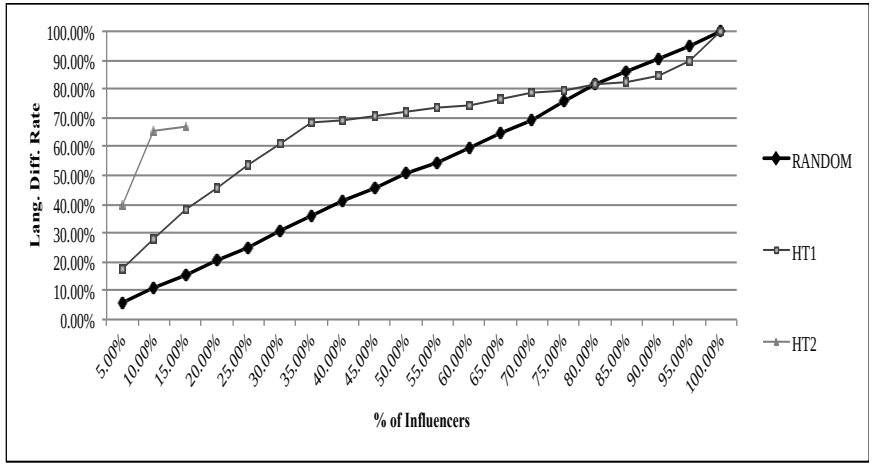
Language Diffusion Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



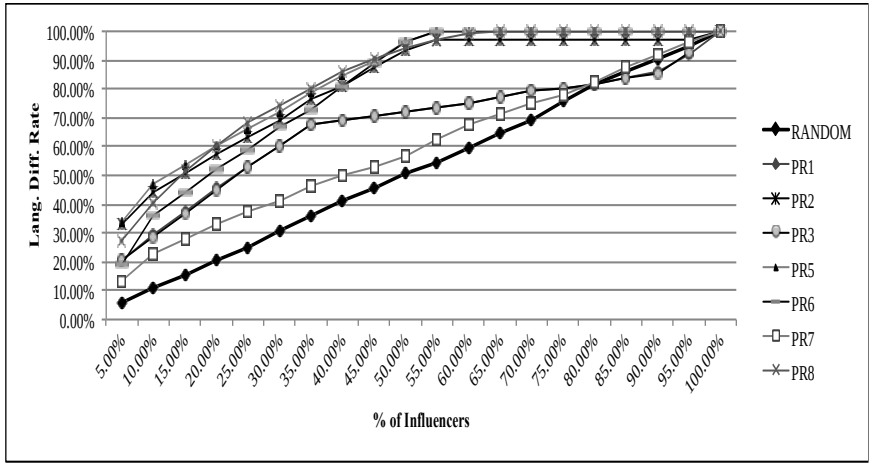
Language Diffusion Rate for Centrality-based Methods: Twitter March Madness Dataset (Window 2, *N*=7,424)



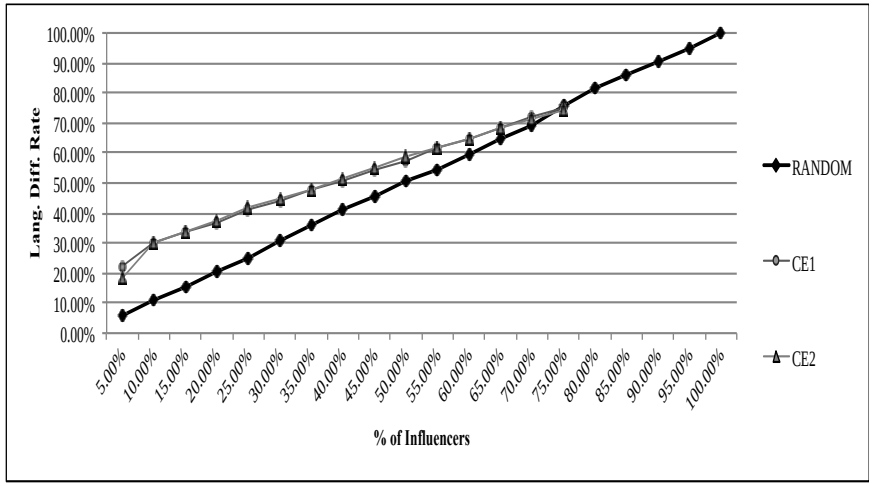
Language Diffusion Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



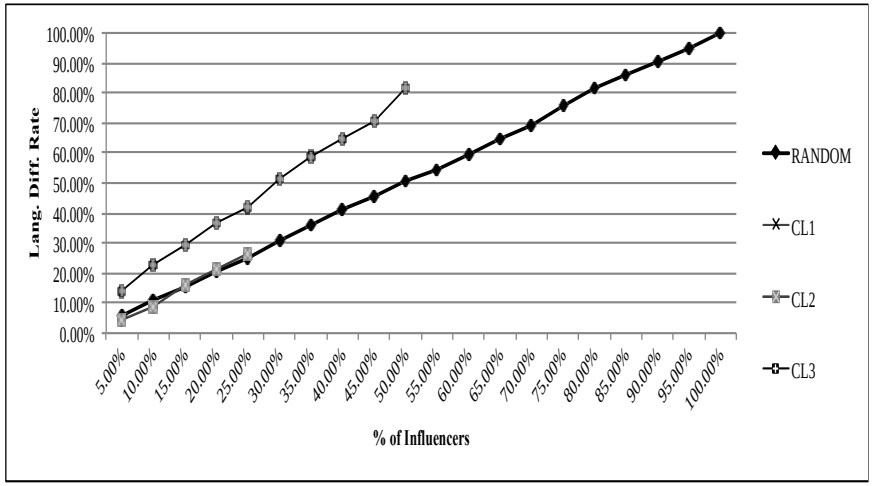
Language Diffusion Rate for HITS-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



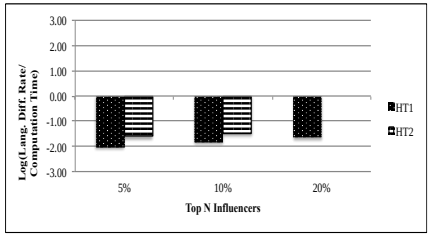
Language Diffusion Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



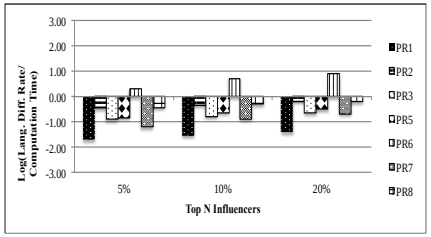
Language Diffusion Rate for Centrality-based Methods: Twitter March Madness Dataset (Window 3, *N*=12,438)



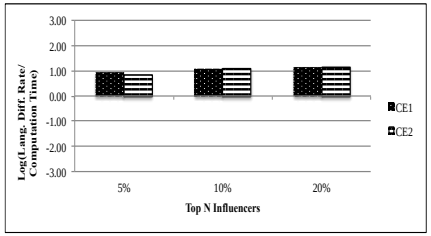
Language Diffusion Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



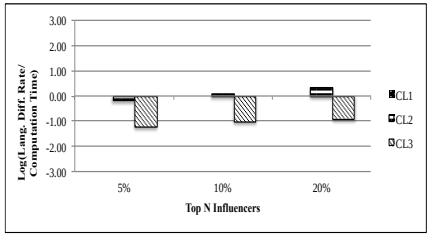
Language Diffusion Rate/ Computation Time for HITS-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



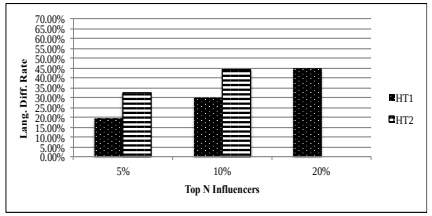
Language Diffusion Rate/ Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



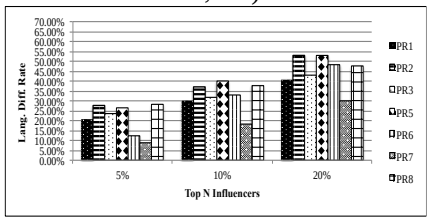
Language Diffusion Rate/ Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 1, *N*=2,039)



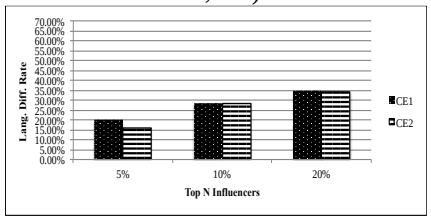
Language Diffusion Rate/ Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



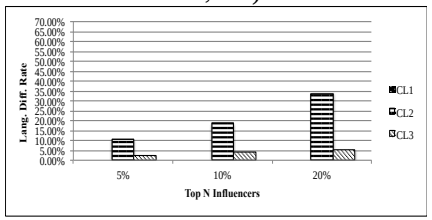
Language Diffusion Rate for HITS-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



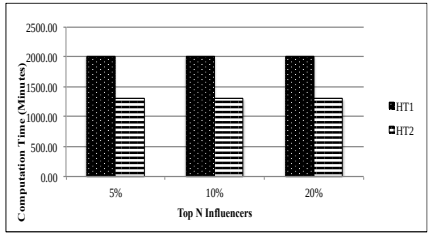
Language Diffusion Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



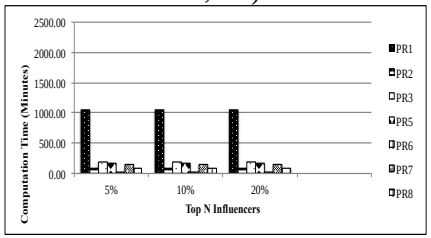
Language Diffusion Rate for Centrality-based Methods: Twitter March Madness Dataset (Window 1, *N*=2,039)



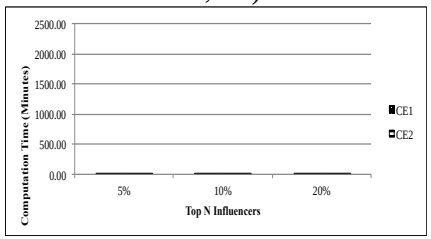
Language Diffusion Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



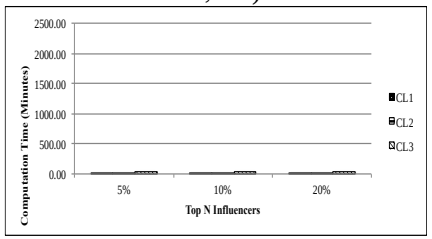
Computation Time for HITS-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



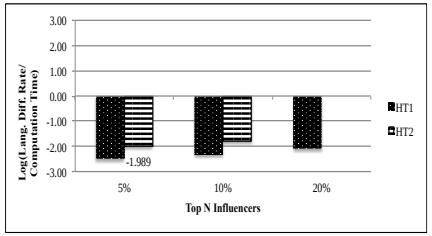
Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



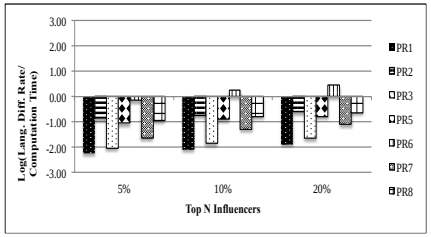
Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 1, *N*=2,039)



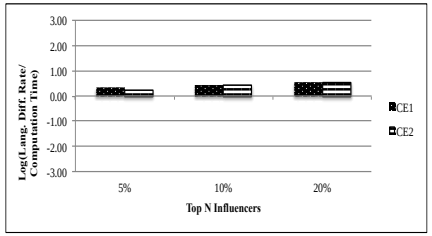
Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



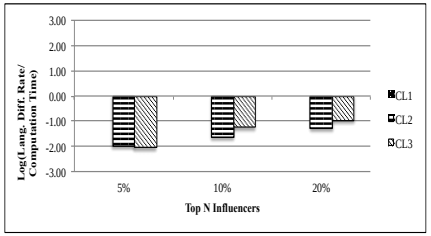
Language Diffusion Rate/ Computation Time for HITS-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



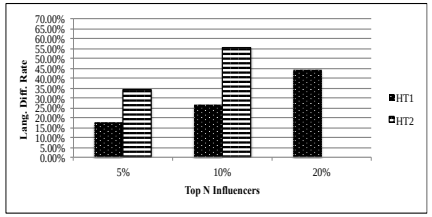
Language Diffusion Rate/ Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



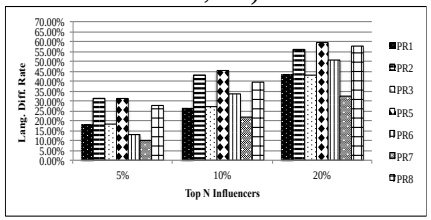
Language Diffusion Rate/ Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 2, *N*=7,424)



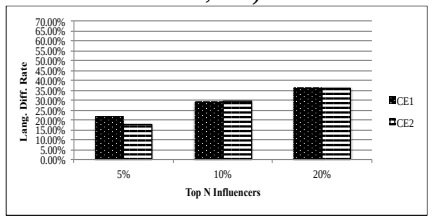
Language Diffusion Rate/ Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



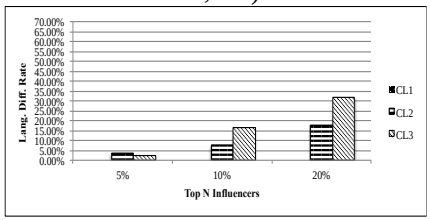
Language Diffusion Rate for HITS -based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



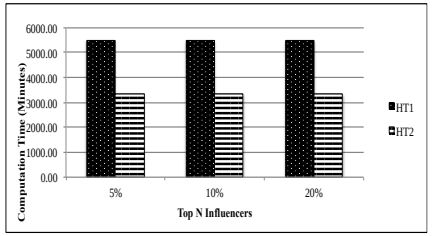
Language Diffusion Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



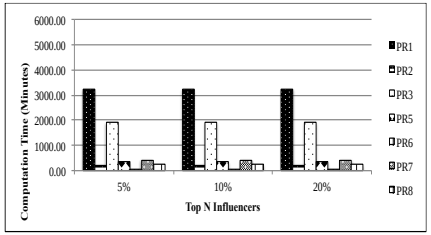
Language Diffusion Rate for Centrality-based Methods: Twitter March Madness Dataset (Window 2, *N*=7,424)



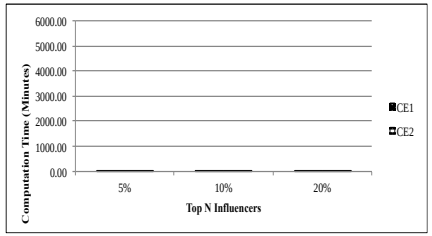
Language Diffusion Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



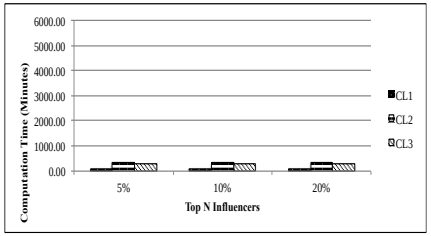
Computation Time for HITS -based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



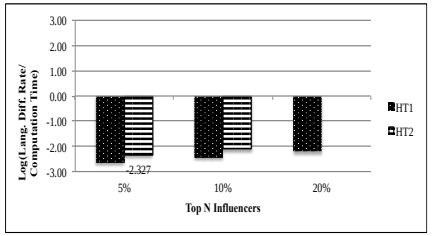
Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



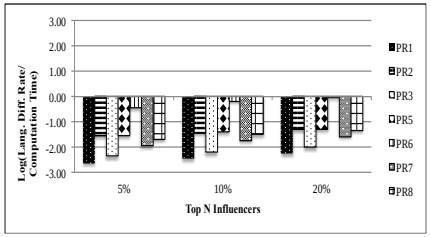
Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 2, *N*=7,424)



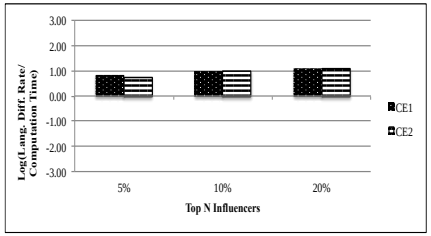
Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



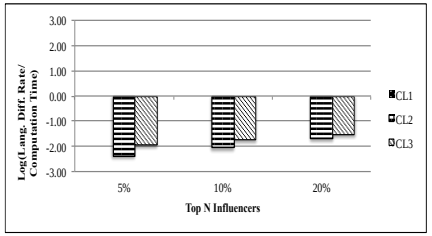
Language Diffusion Rate/ Computation Time for HITS-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



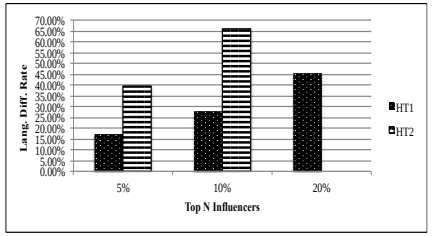
Language Diffusion Rate/ Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



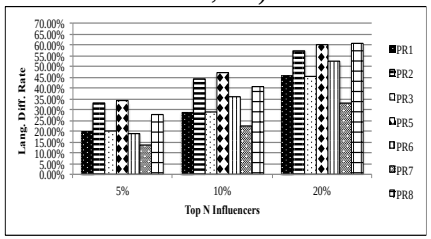
Language Diffusion Rate/ Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 3, *N*=12,438)



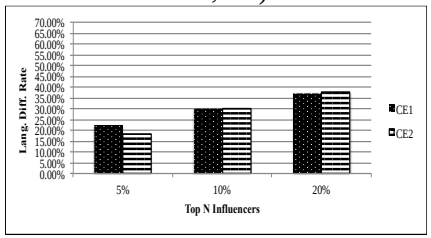
Language Diffusion Rate/ Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



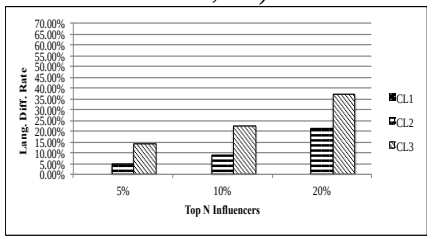
Language Diffusion Rate for HITS -based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



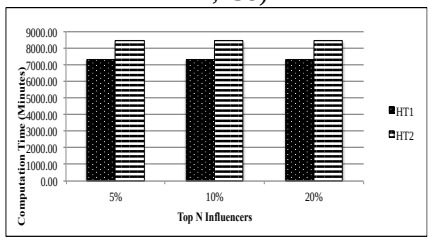
Language Diffusion Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438))



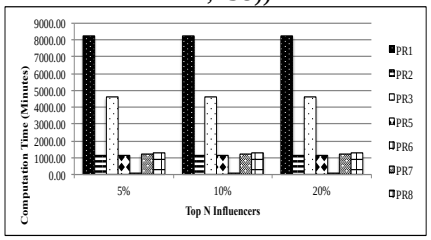
Language Diffusion Rate for Centrality-based Methods: Twitter March Madness Dataset (Window 3, *N*=12,438)



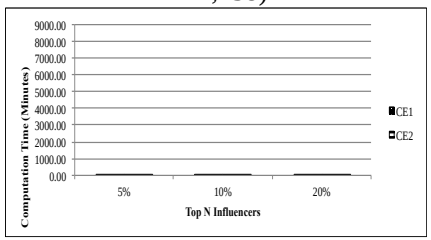
Language Diffusion Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



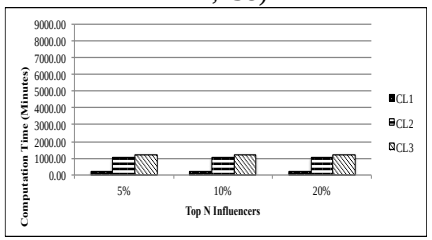
Computation Time for HITS -based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



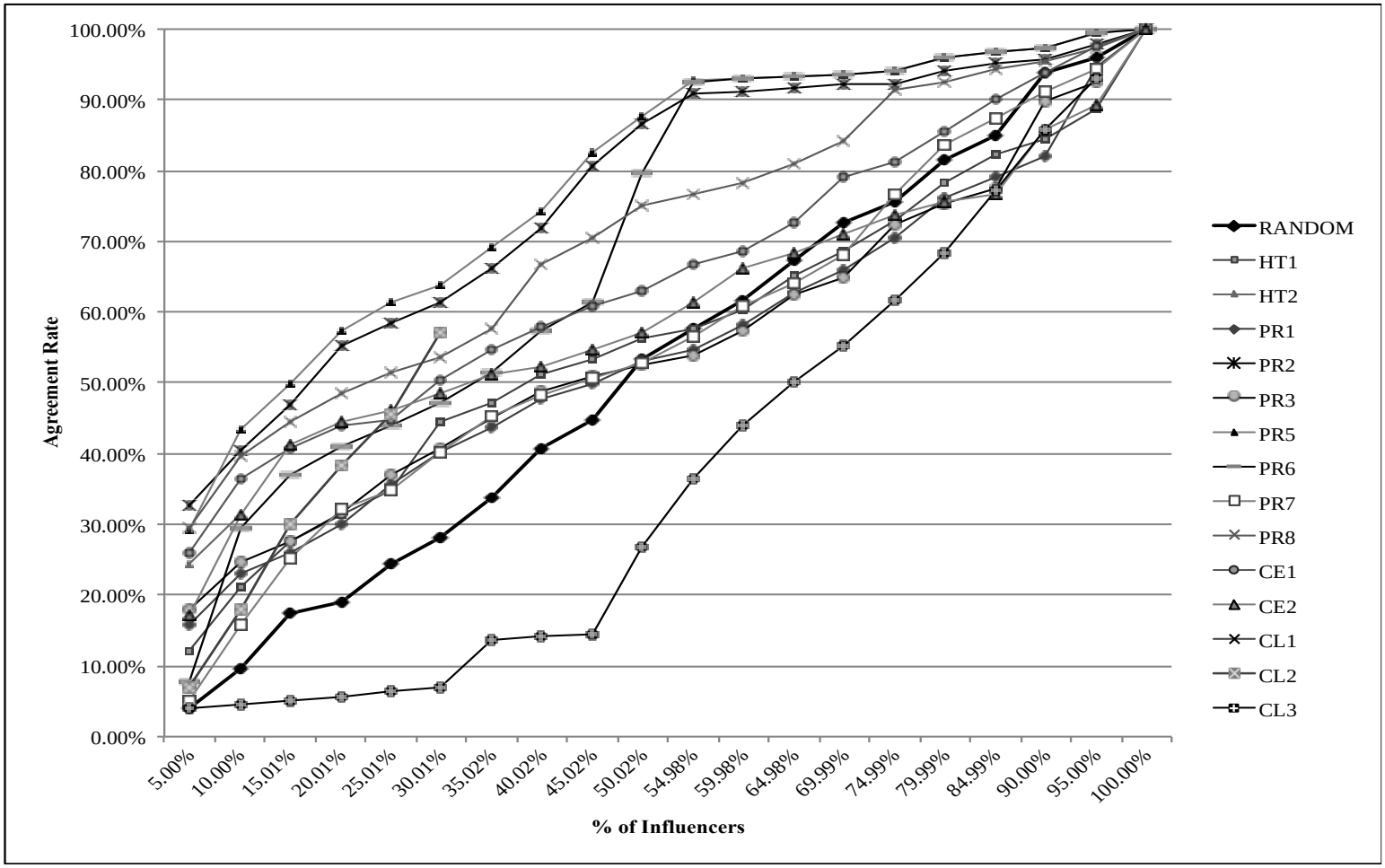
Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



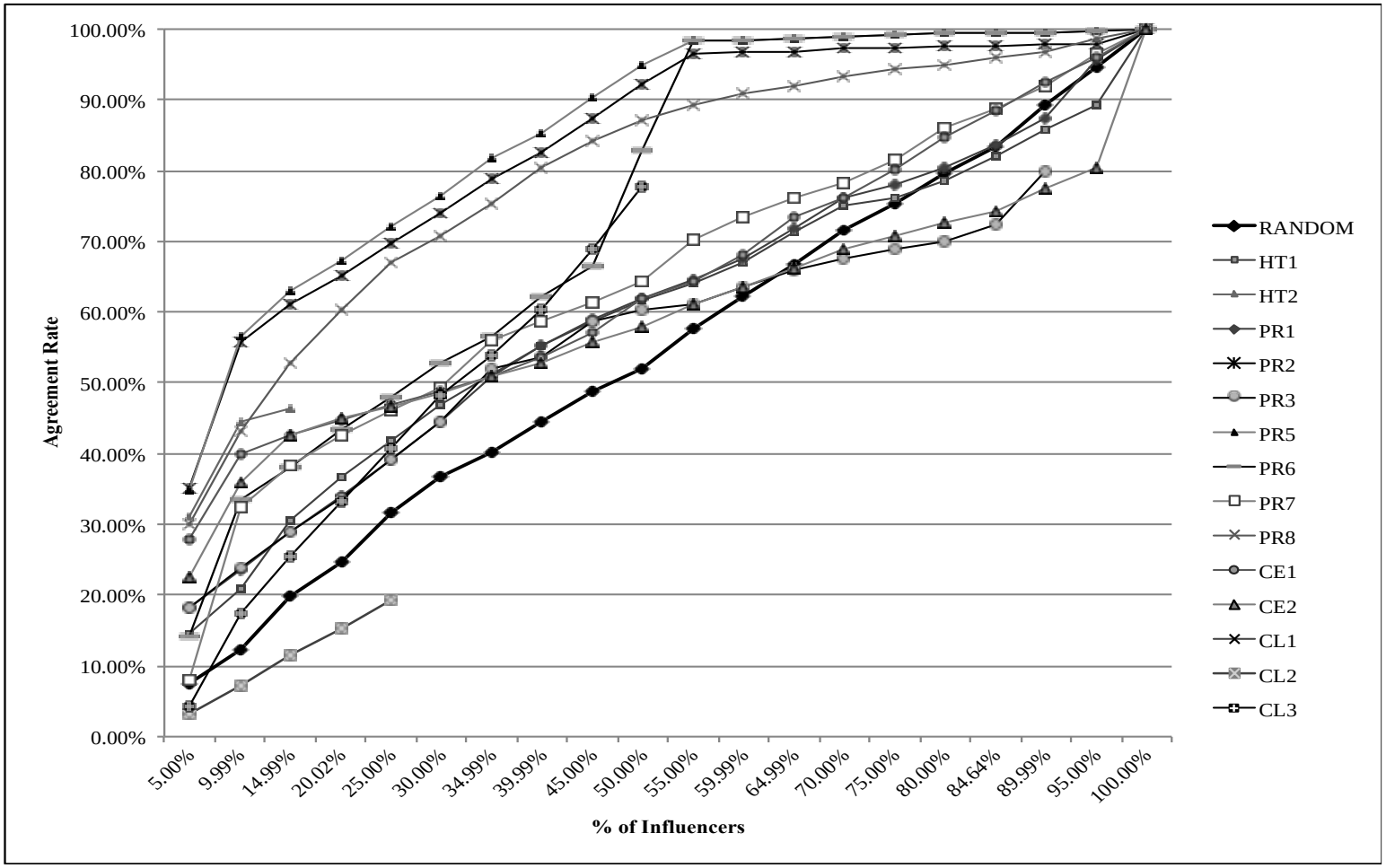
Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 3, *N*=12,438)



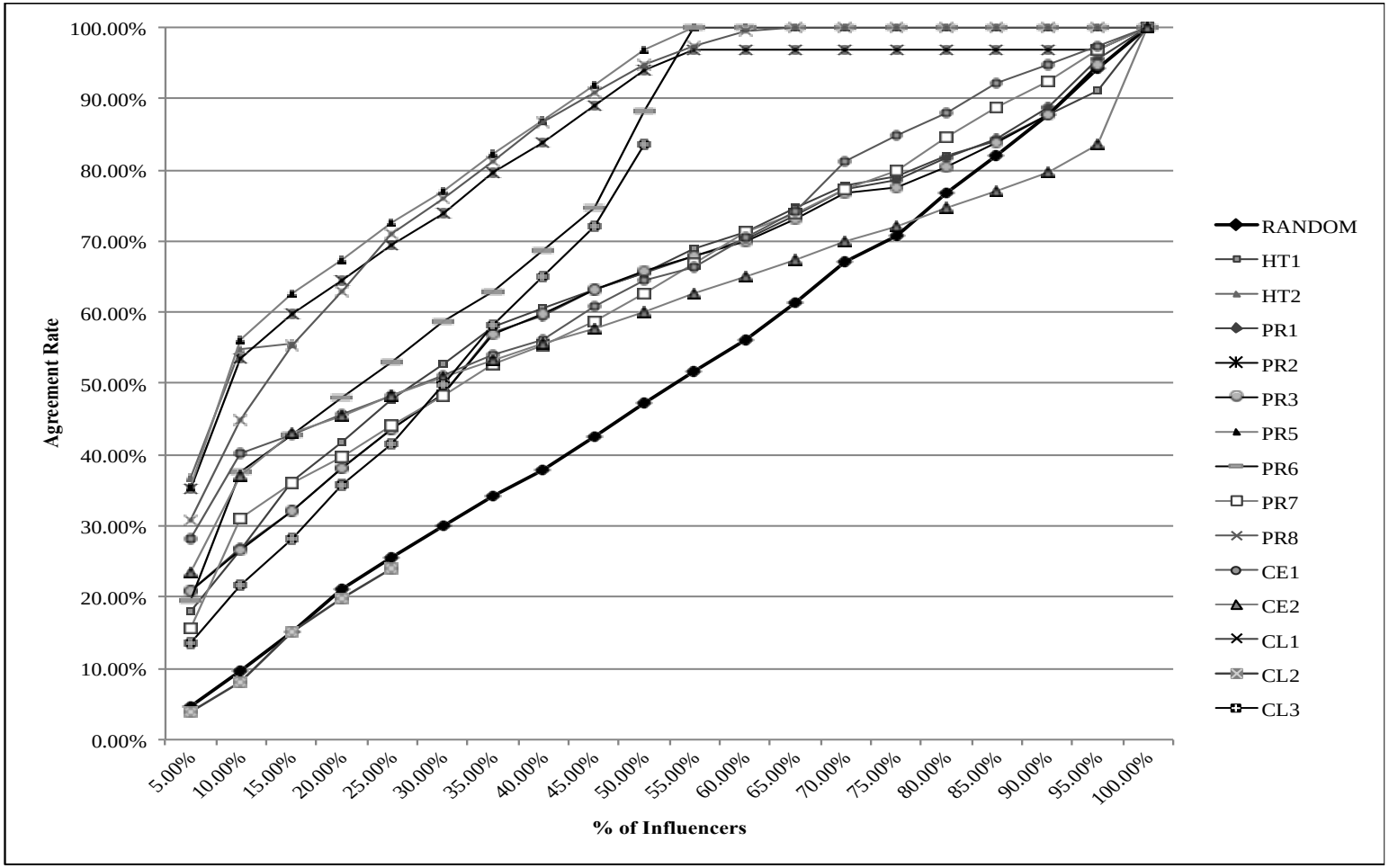
Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



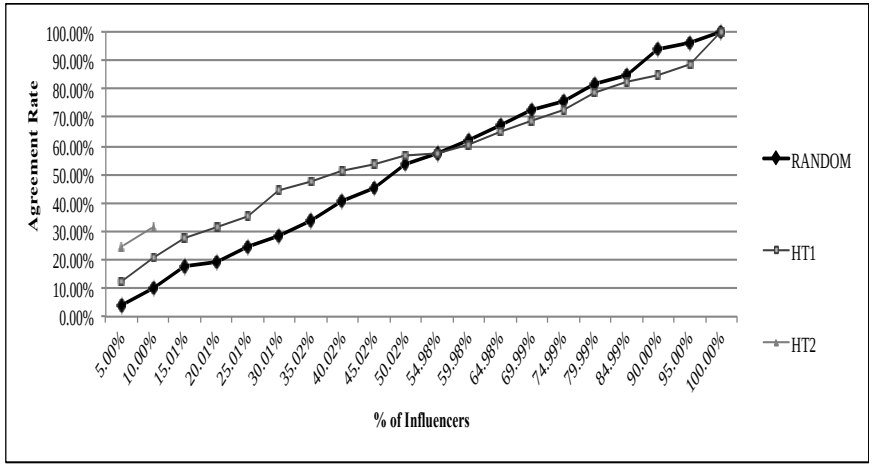
Agreement Rate for Different Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



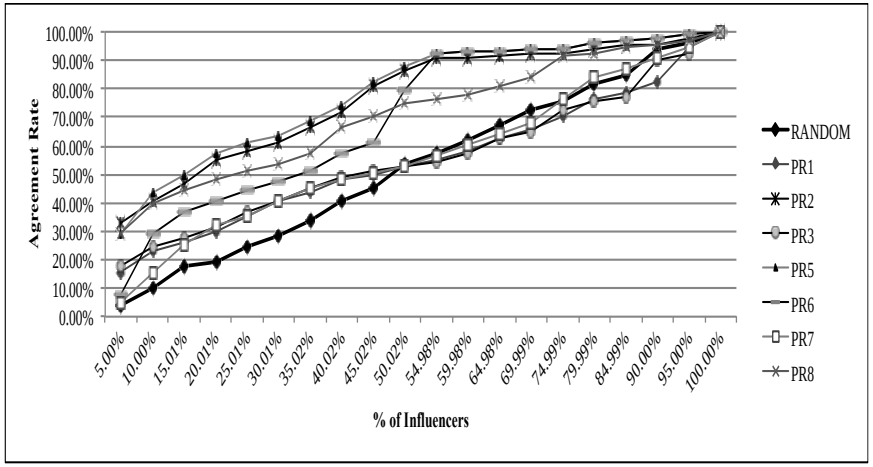
Agreement Rate for Different Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



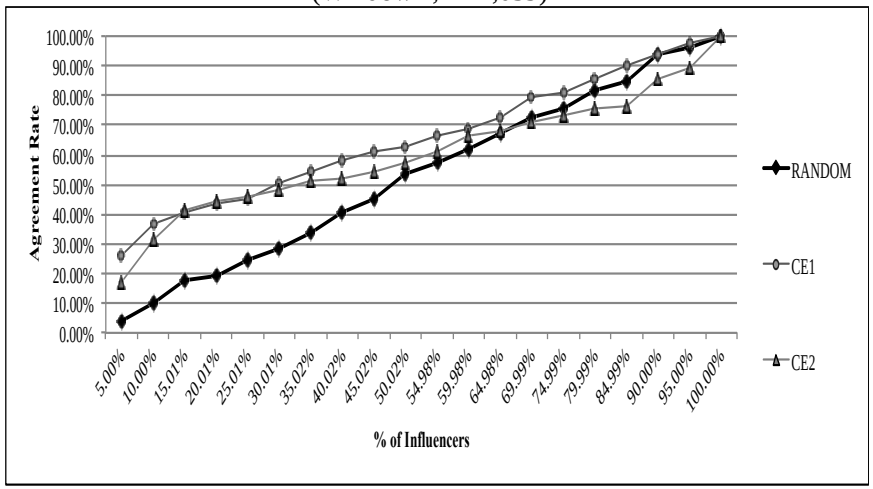
Agreement Rate for Different Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



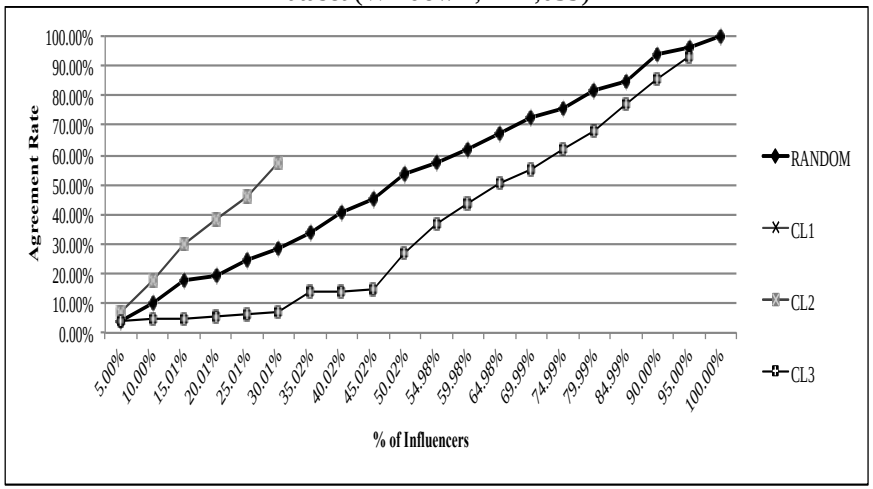
Agreement Rate for HITS-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



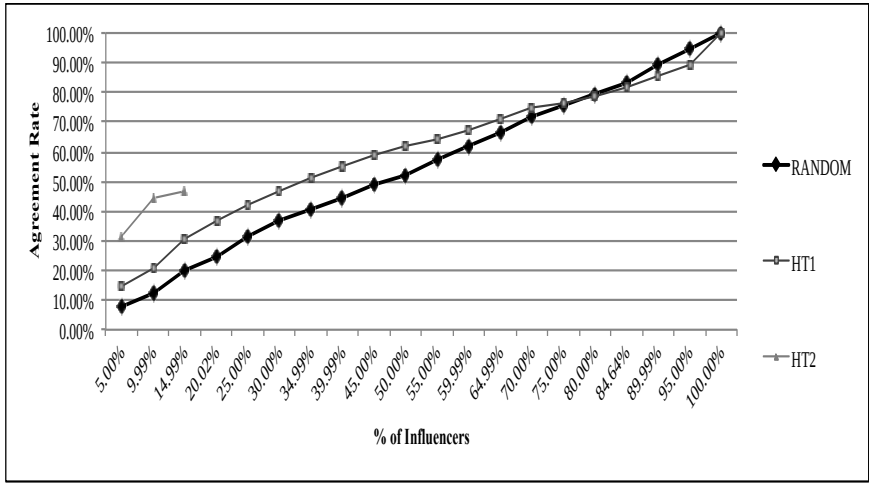
Agreement Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



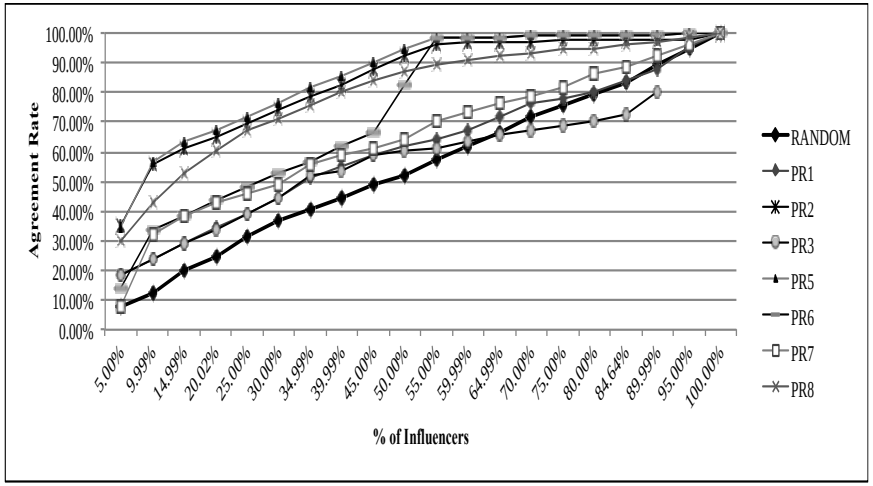
Agreement Rate for Centrality-based Methods: Twitter March Madness Dataset (Window 1, *N*=2,039)



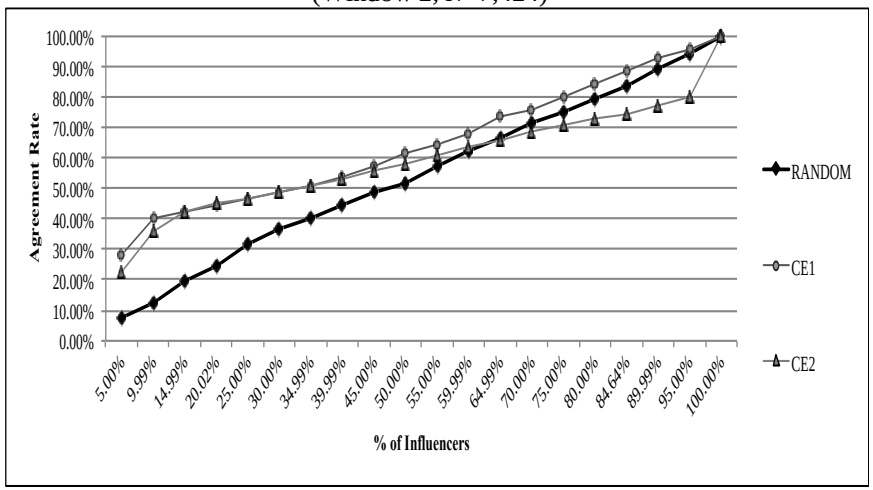
Agreement Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



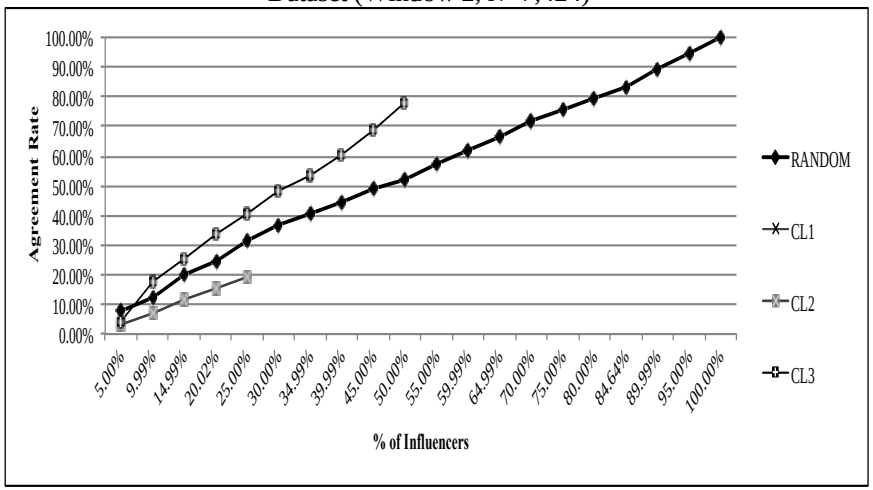
Agreement Rate for HITS-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



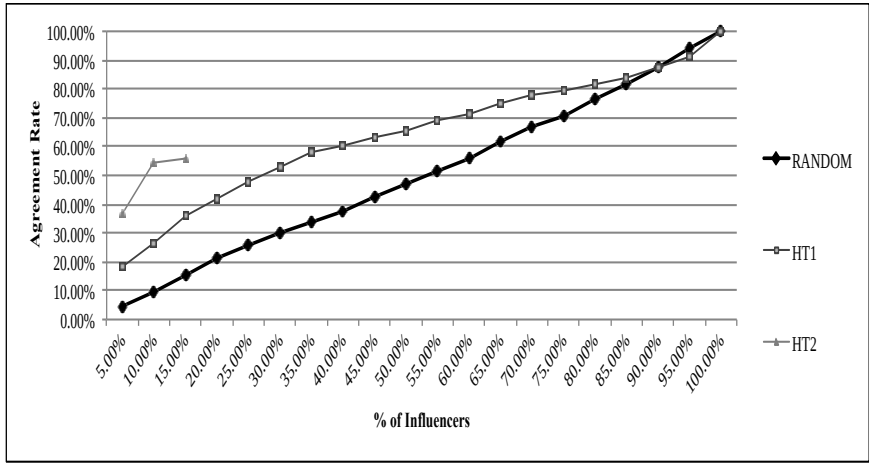
Agreement Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



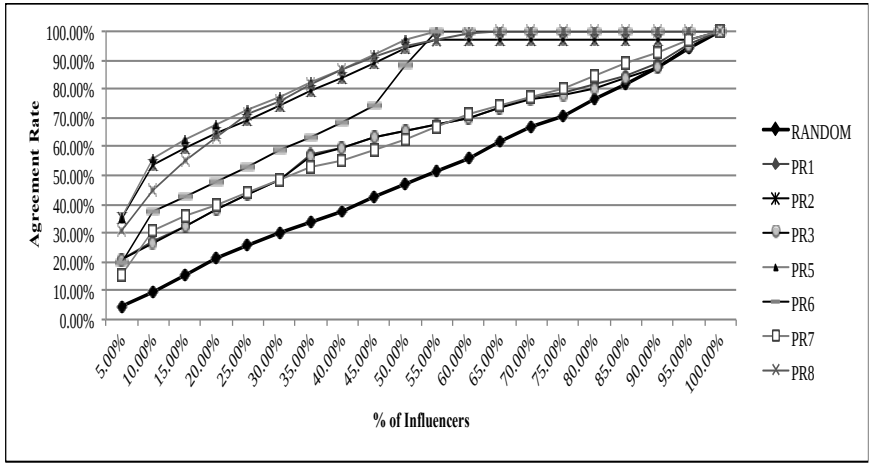
Agreement Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



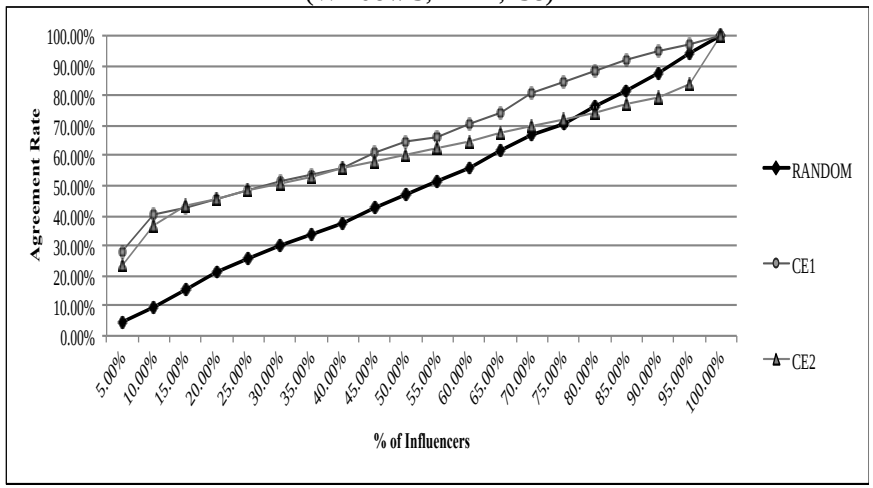
Agreement Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



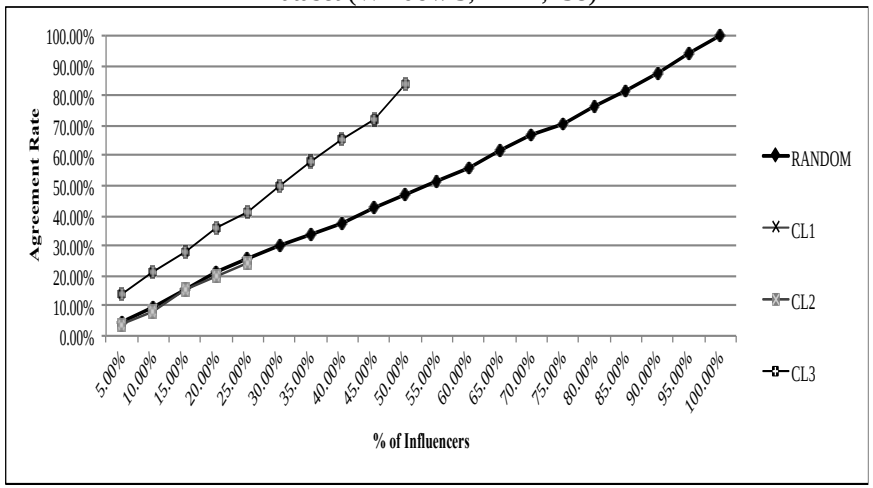
Agreement Rate for HITS-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



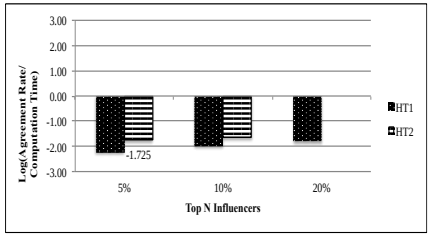
Agreement Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



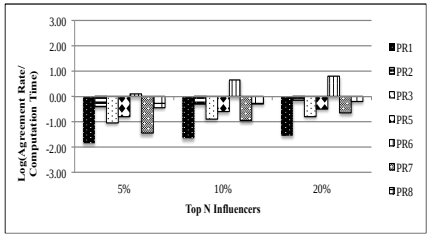
Agreement Rate for Centrality-based Methods: Twitter March Madness Dataset (Window 3, *N*=12,438)



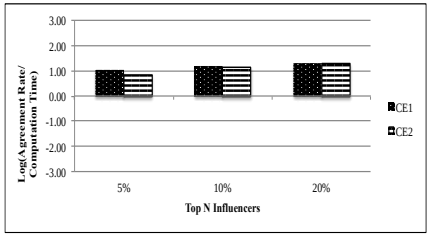
Agreement Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



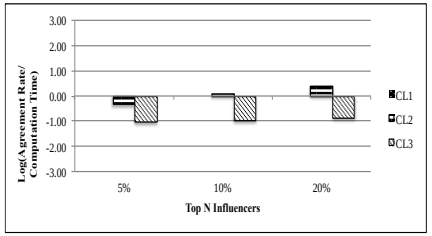
Agreement Rate/ Computation Time for HITS-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



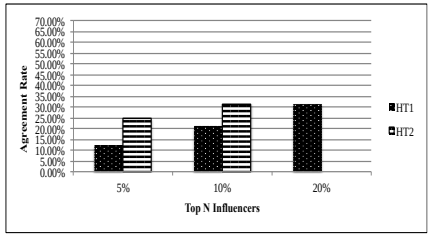
Agreement Rate/ Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



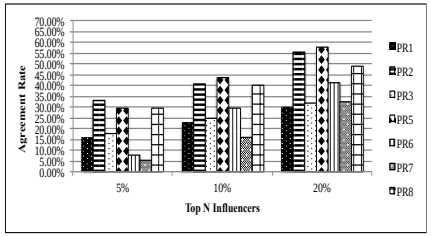
Agreement Rate/ Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 1, *N*=2,039)



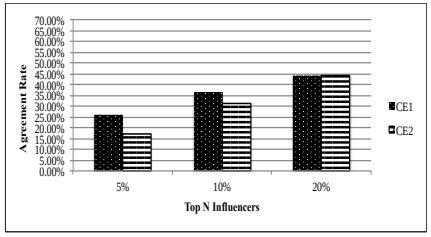
Agreement Rate/ Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



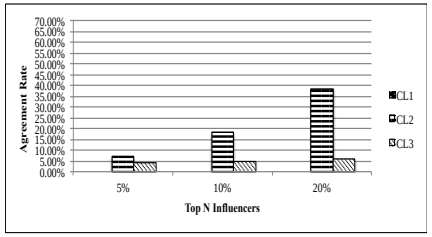
Agreement Rate for HITS-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



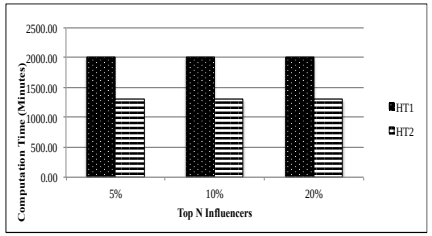
Agreement Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



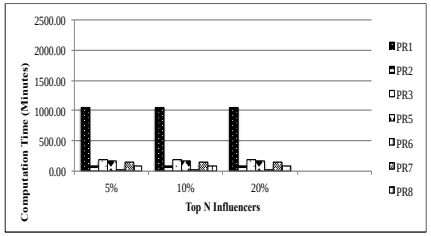
Agreement Rate for Centrality-based Methods: Twitter March Madness Dataset (Window 1, *N*=2,039)



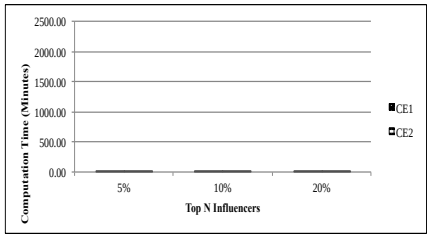
Agreement Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



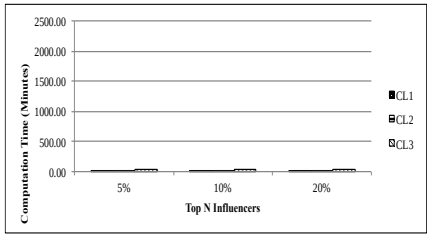
Computation Time for HITS-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



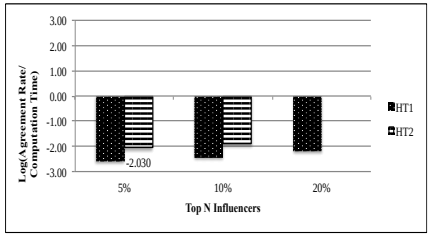
Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



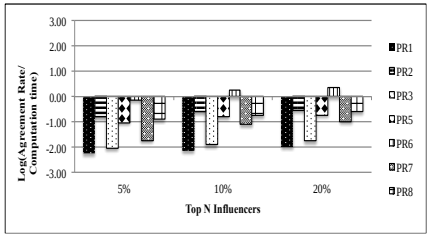
Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 1, *N*=2,039)



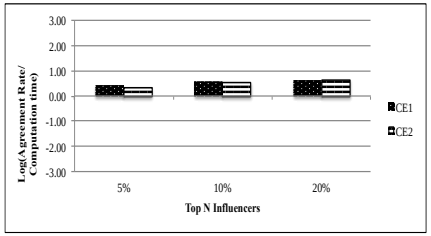
Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 1, *N*=2,039)



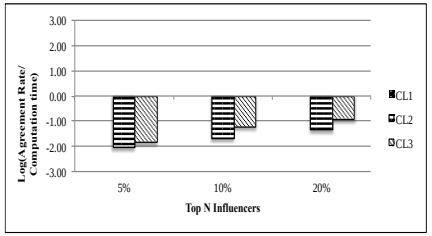
Agreement Rate/ Computation Time for HITS-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



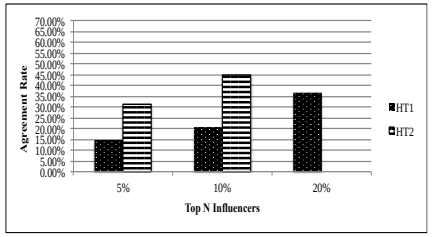
Agreement Rate/ Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



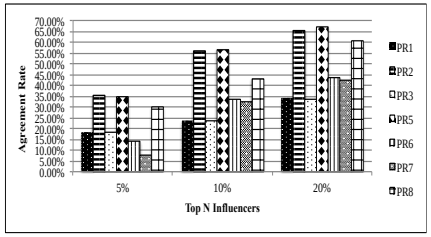
Agreement Rate/ Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 2, *N*=7,424)



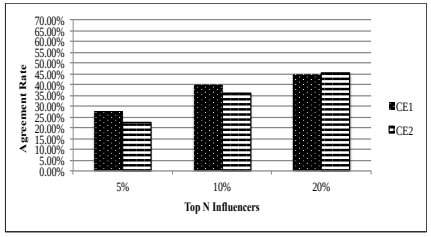
Agreement Rate/ Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



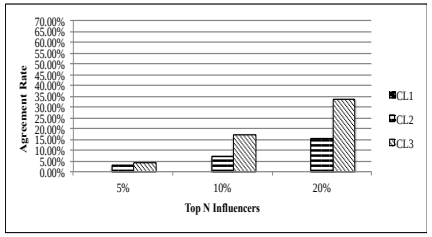
Agreement Rate for HITS -based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



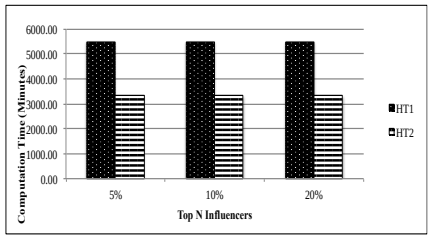
Agreement Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



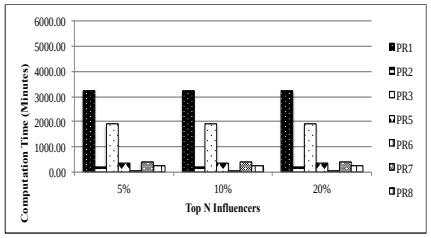
Agreement Rate for Centrality-based Methods: Twitter March Madness Dataset (Window 2, *N*=7,424)



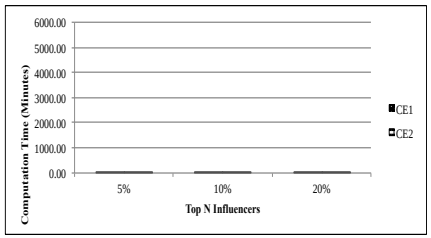
Agreement Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



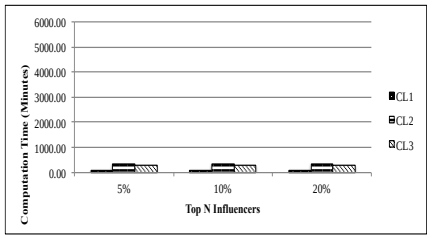
Computation Time for HITS -based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



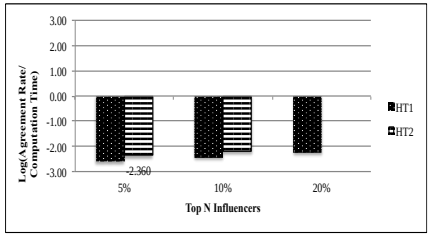
Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



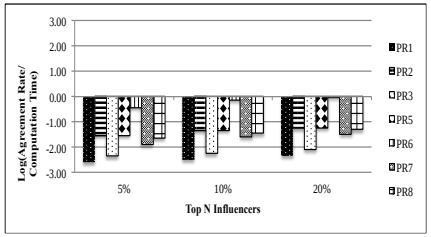
Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 2, *N*=7,424)



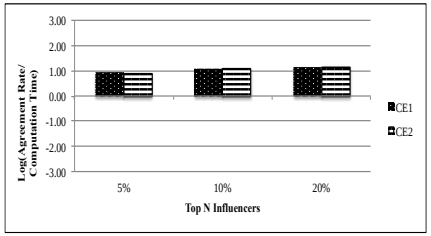
Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 2, *N*=7,424)



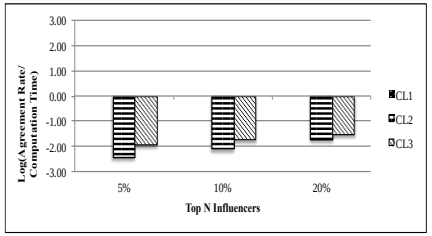
Agreement Rate/ Computation Time for HITS-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



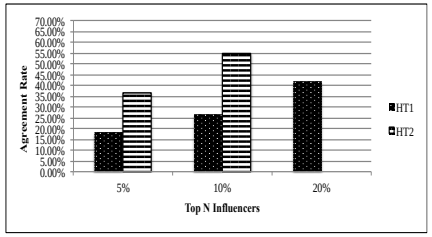
Agreement Rate/ Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



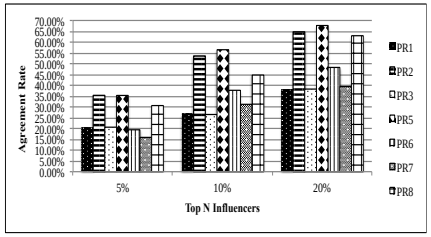
Agreement Rate/ Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 3, *N*=12,438)



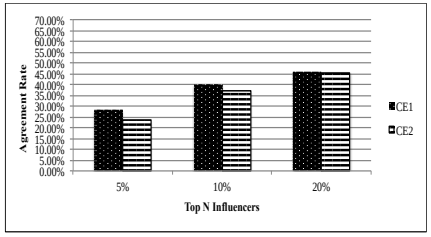
Agreement Rate/ Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



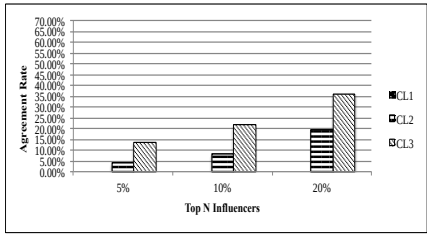
Agreement Rate for HITS -based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



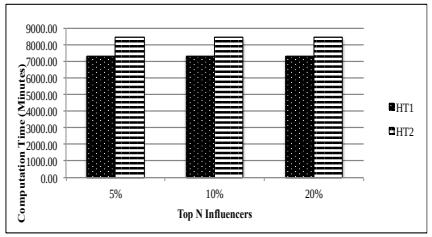
Agreement Rate for PageRank-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438))



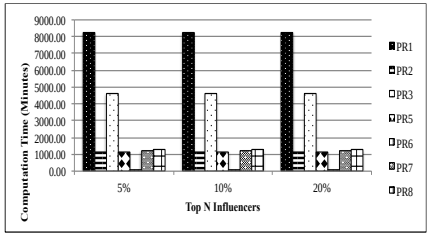
Agreement Rate for Centrality-based Methods: Twitter March Madness Dataset (Window 3, *N*=12,438)



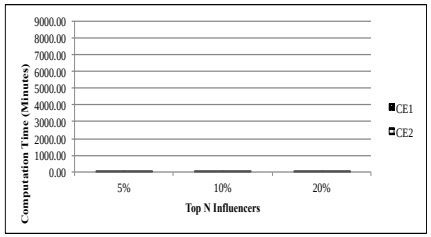
Agreement Rate for Clustering-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



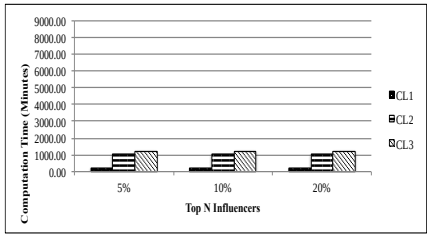
Computation Time for HITS -based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)



Computation Time for PageRank-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)

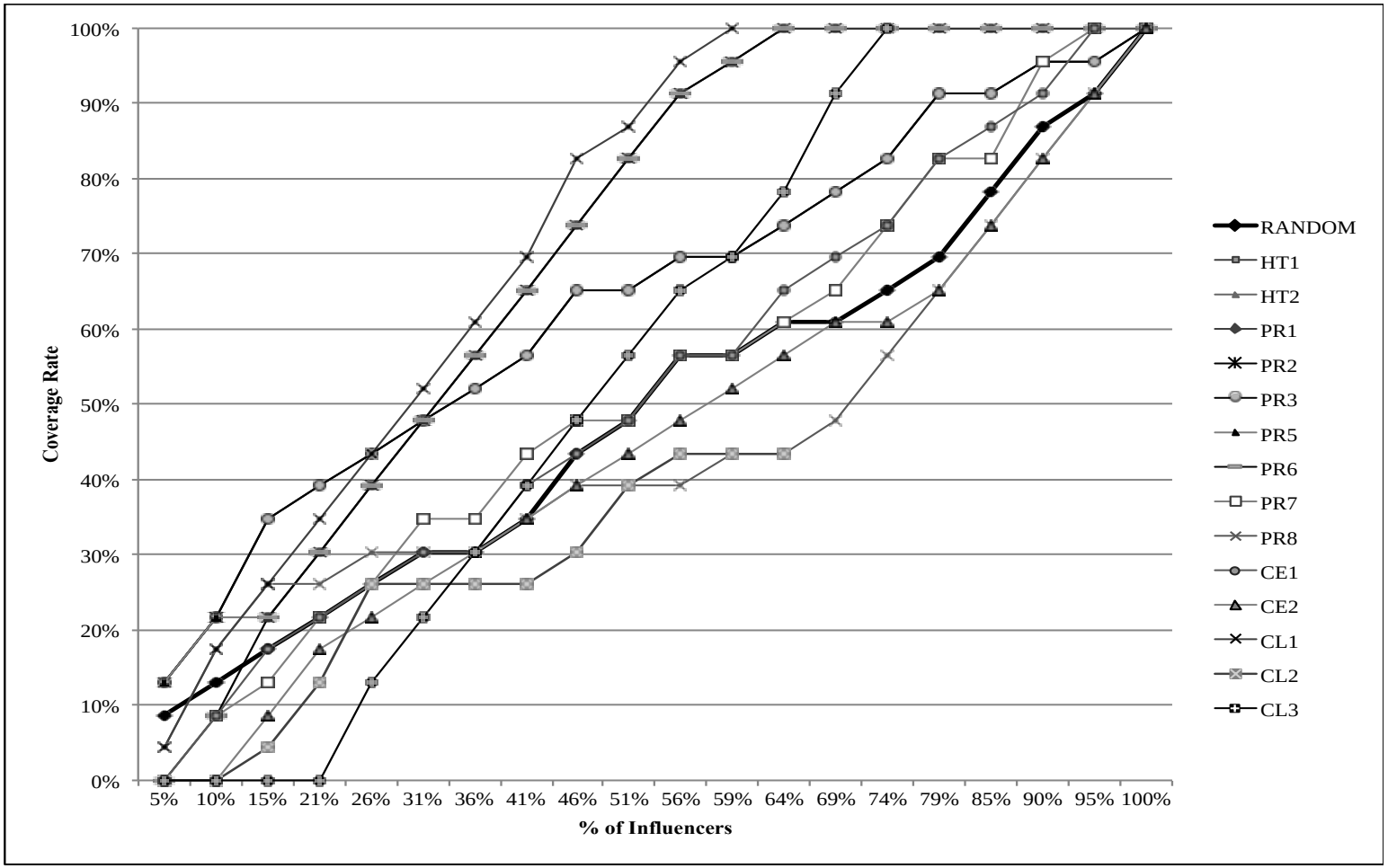


Computation Time for Centrality-based Methods: Twitter March Madness Dataset (Window 3, *N*=12,438)

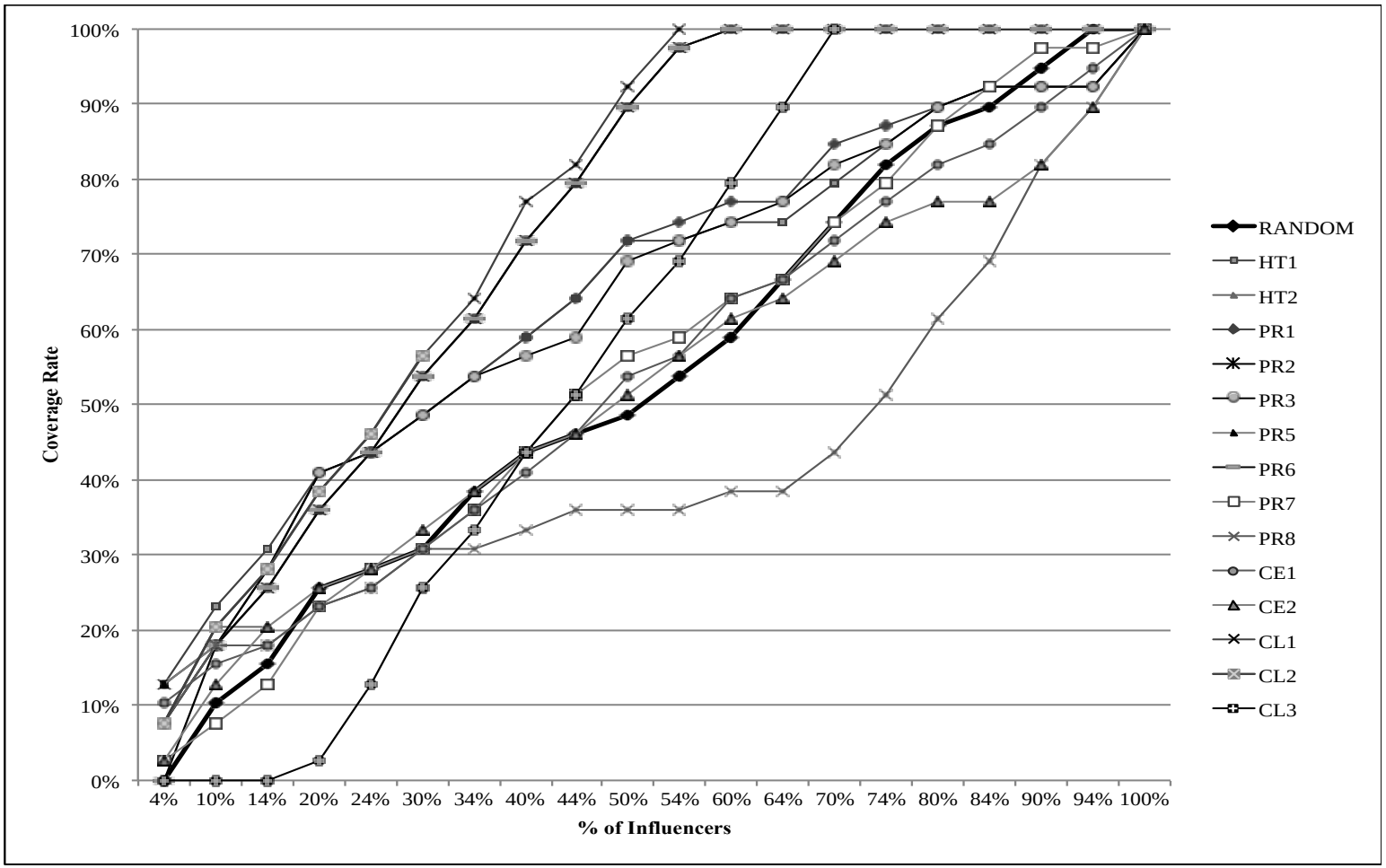


Computation Time for Clustering-based Algorithms: Twitter March Madness Dataset (Window 3, *N*=12,438)

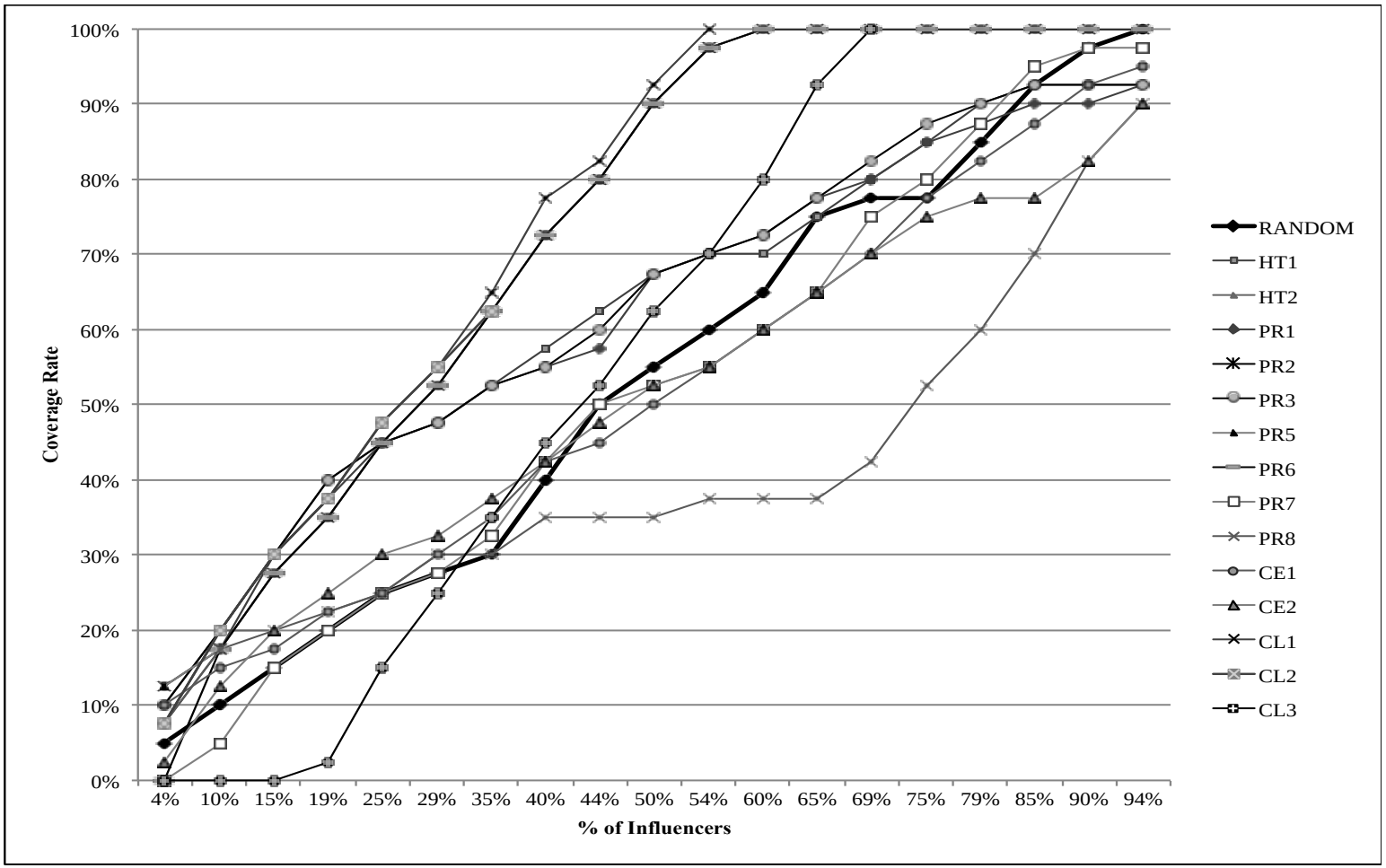
Appendix IV: Experiment Results from the Twitter KY Derby Dataset



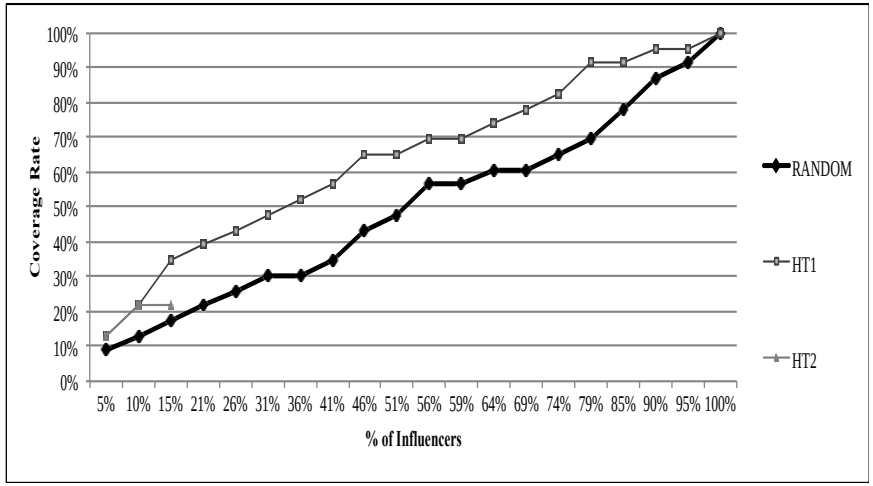
Coverage Rate for Different Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



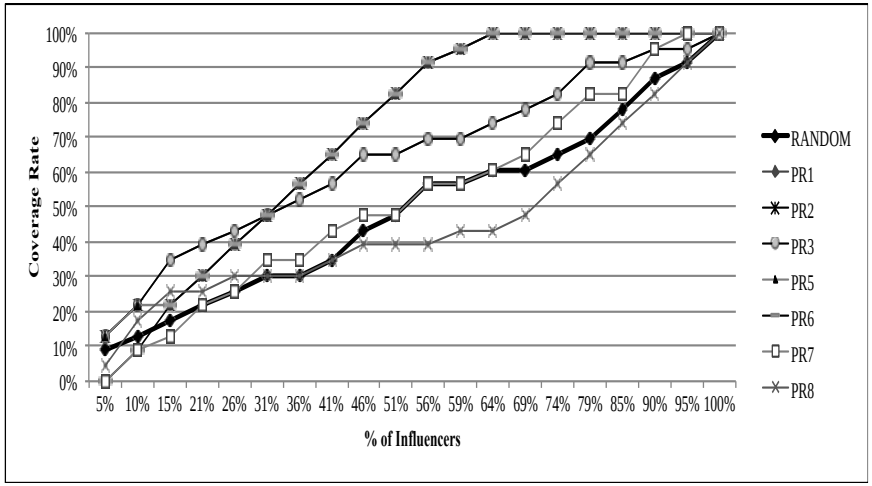
Coverage Rate for Different Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



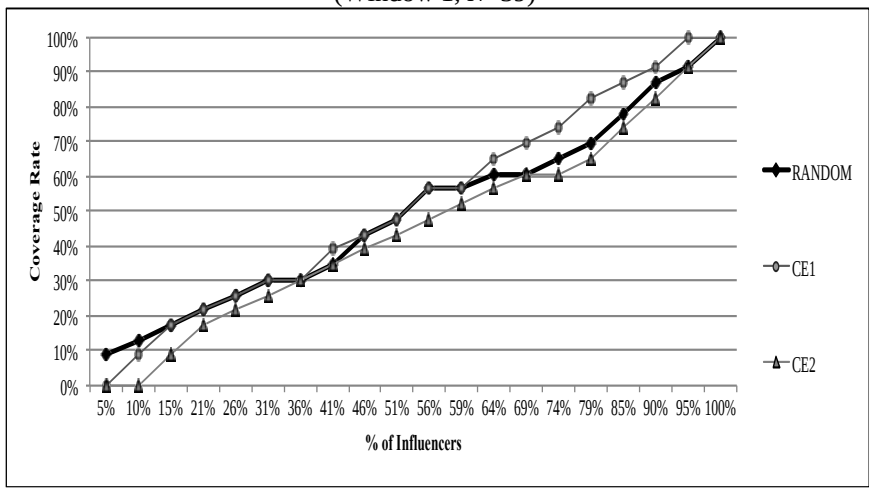
Coverage Rate for Different Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



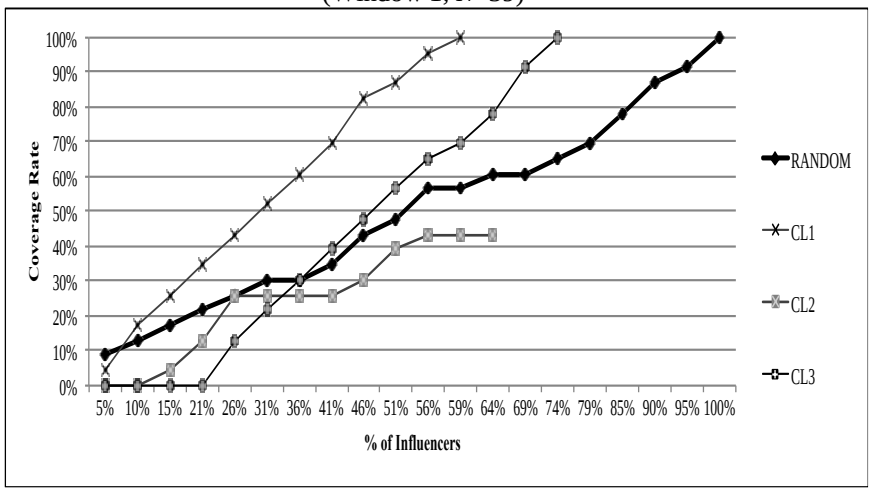
Coverage Rate for HITS-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



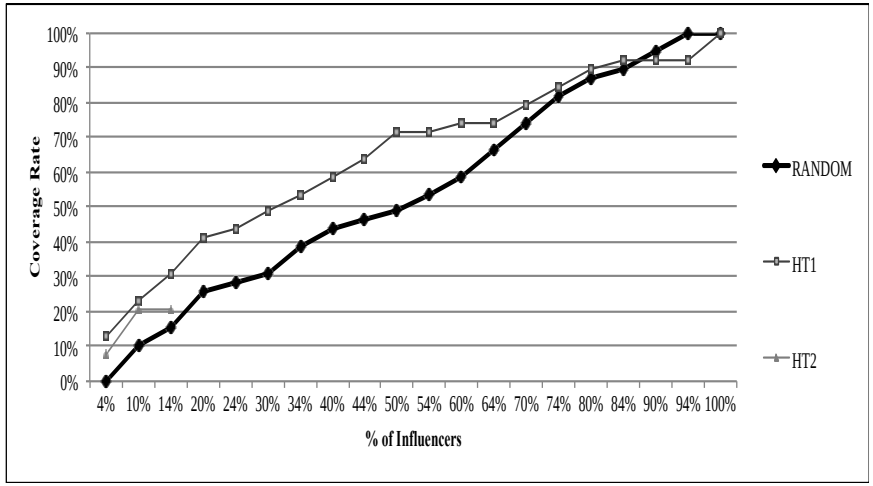
Coverage Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



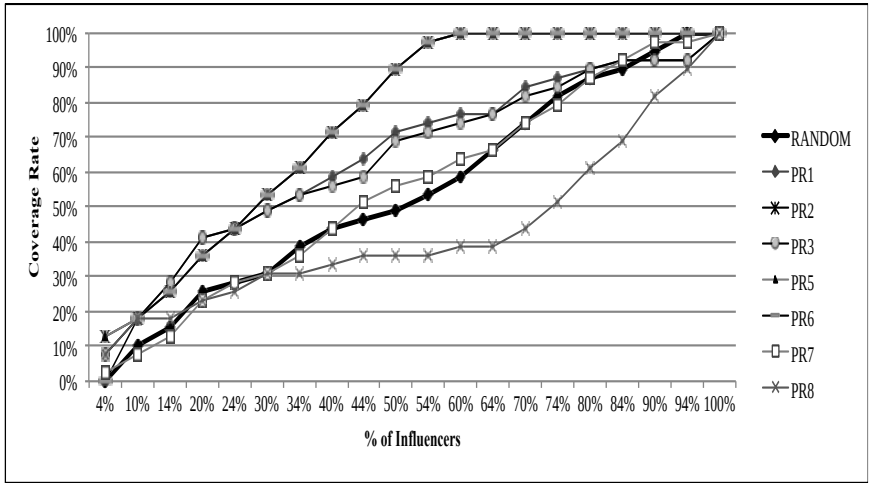
Coverage Rate for Centrality_based Methods: Twitter KY Derby Dataset (Window 1, *N*=39)



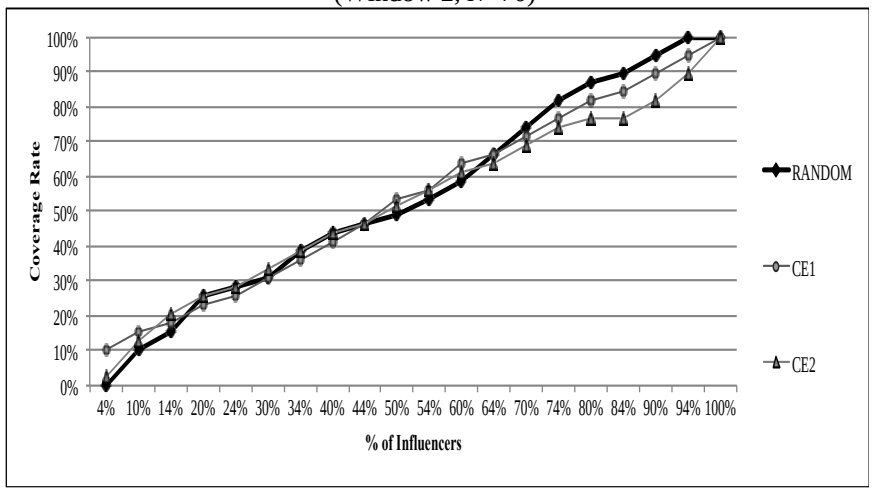
Coverage Rate for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



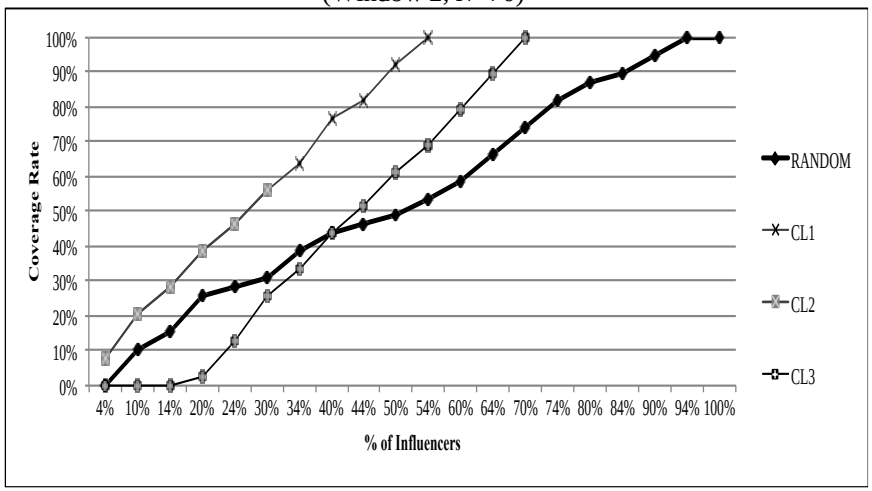
Coverage Rate for HITS-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



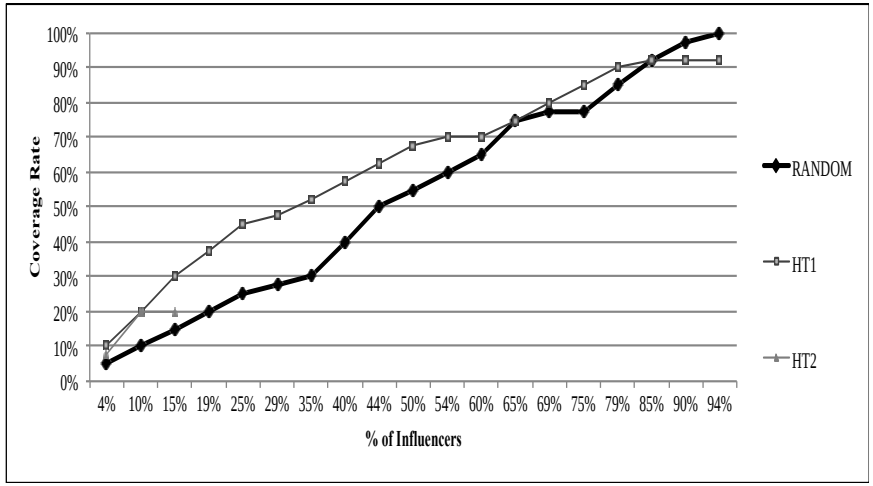
Coverage Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



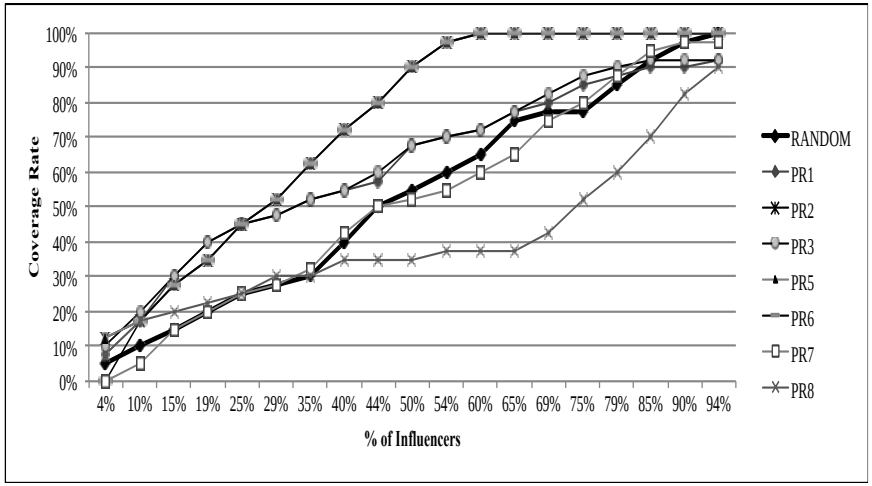
Coverage Rate for Centrality_based Methods: Twitter KY Derby Dataset (Window 2, *N*=70)



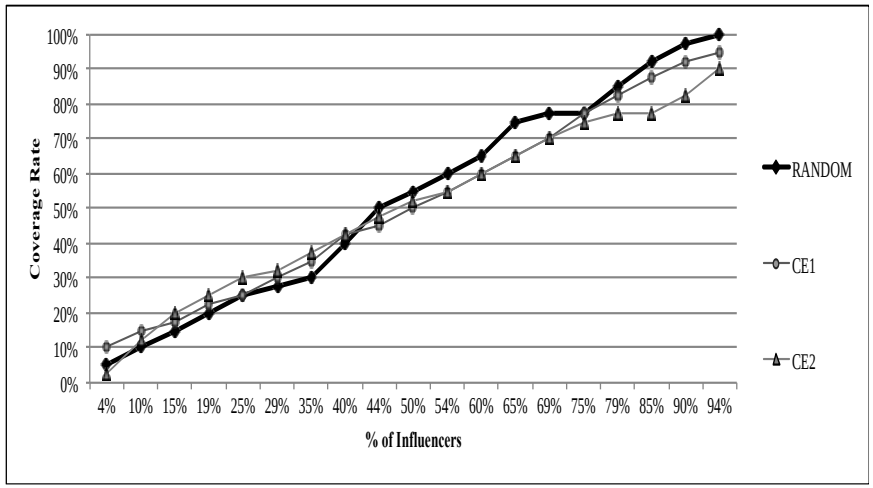
Coverage Rate for Clustering -based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



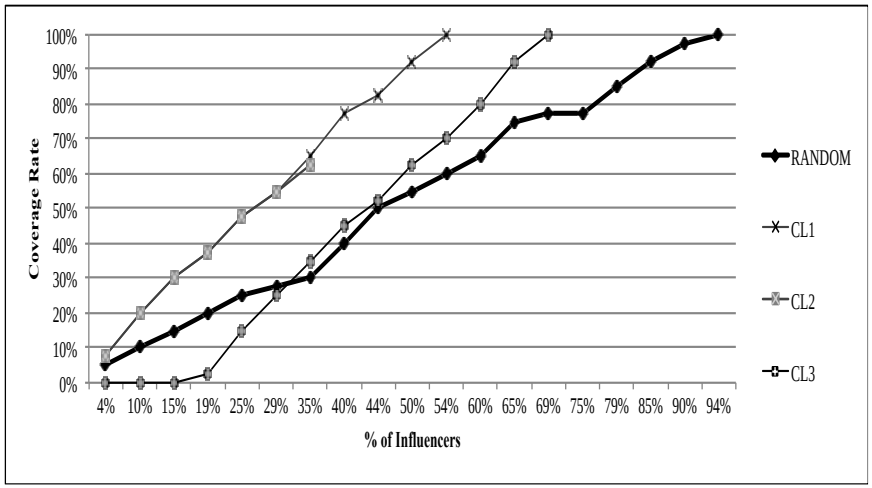
Coverage Rate for HITS-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



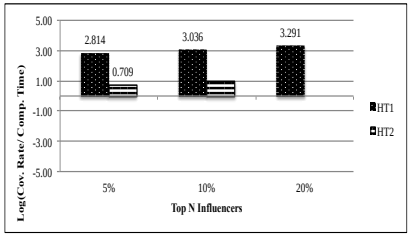
Coverage Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



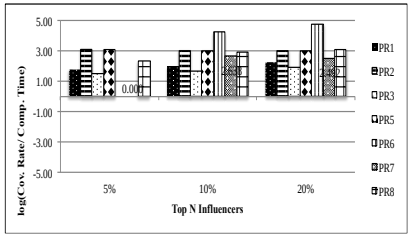
Coverage Rate for Centrality_based Methods: Twitter KY Derby Dataset (Window 3, *N*=72)



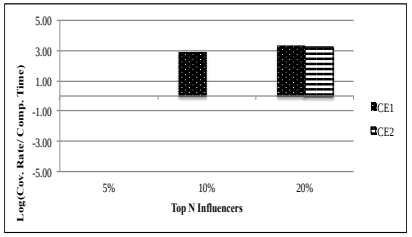
Coverage Rate for Clustering -based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



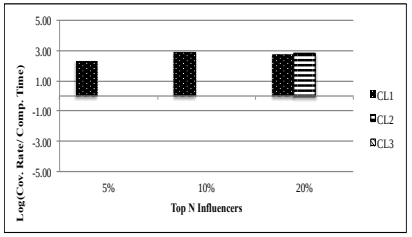
Coverage Rate/ Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 1, N=39)



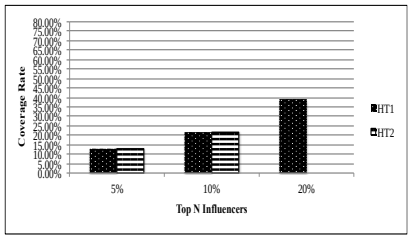
Coverage Rate/ Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 1, N=39)



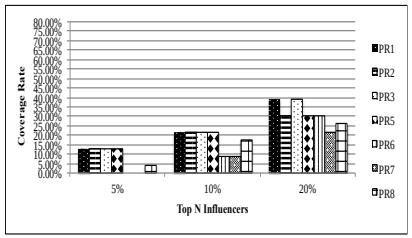
Coverage Rate/ Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 1, N=39)



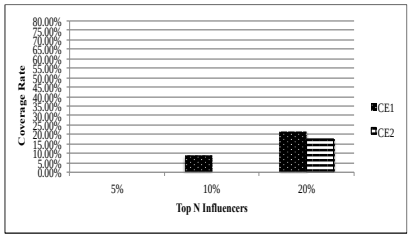
Coverage Rate/ Computation Time for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 1, N=39)



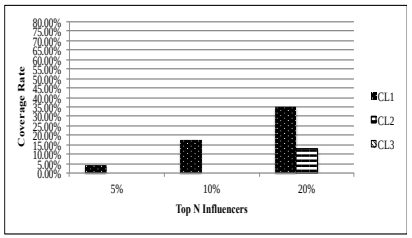
Coverage Rate for HITS-based Algorithms: Twitter KY Derby Dataset t (Window 1, N=39)



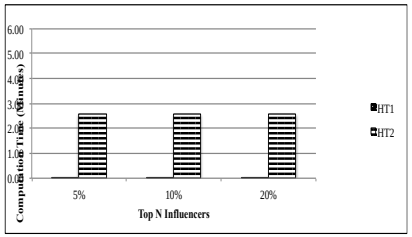
Coverage Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 1, N=39)



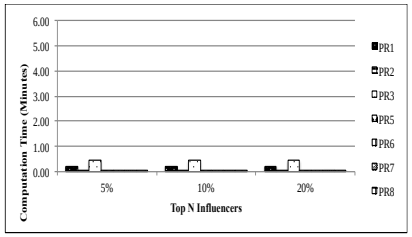
Coverage Rate for Centrality-based Methods: Twitter KY Derby Dataset (Window 1, N=39)



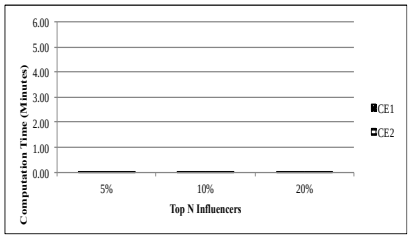
Coverage Rate for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 1, N=39)



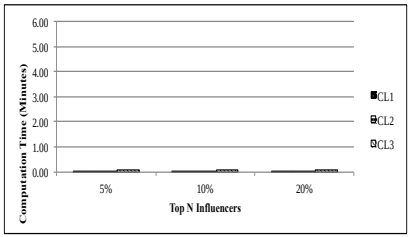
Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 1, N=39)



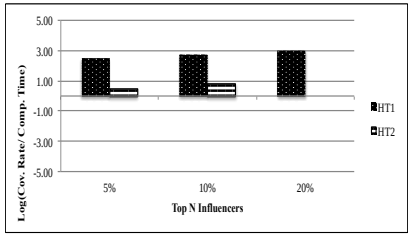
Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 1, N=39)



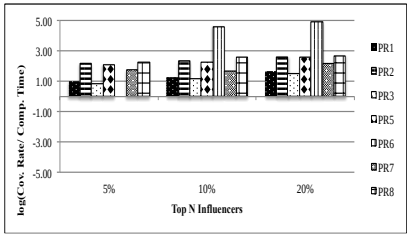
Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 1, N=39)



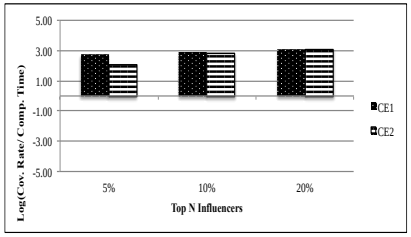
Computation Time for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 1, N=39)



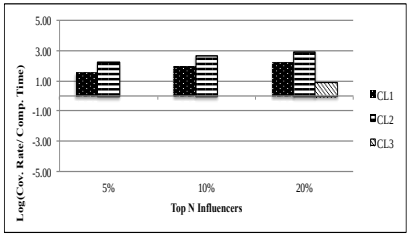
Coverage Rate/ Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



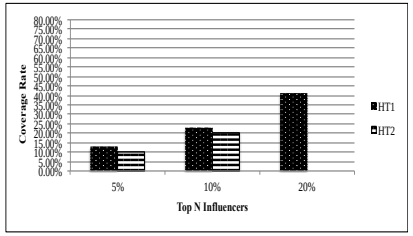
Coverage Rate/ Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



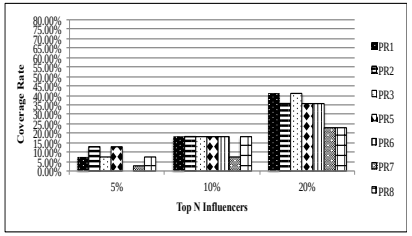
Coverage Rate/ Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 2, *N*=70)



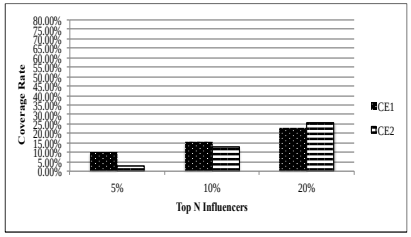
Coverage Rate/ Computation Time for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



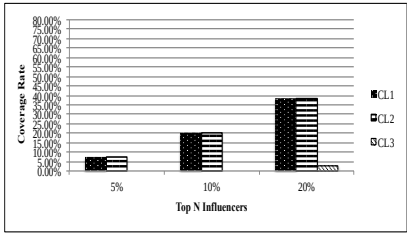
Coverage Rate for HITS-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



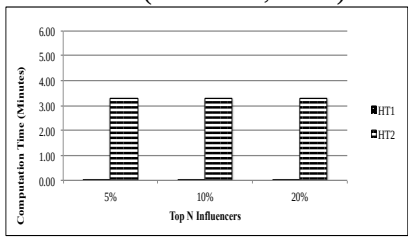
Coverage Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



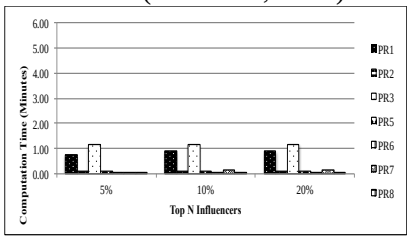
Coverage Rate for Centrality-based Methods: Twitter KY Derby Dataset (Window 2, *N*=70)



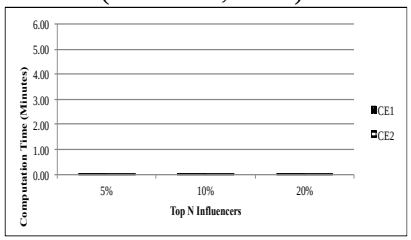
Coverage Rate for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



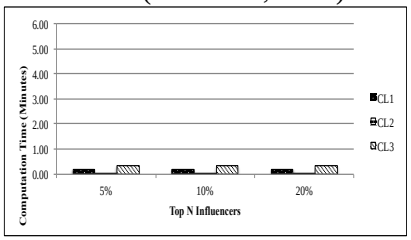
Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



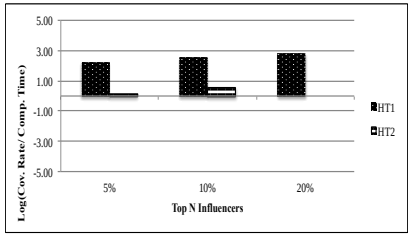
Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



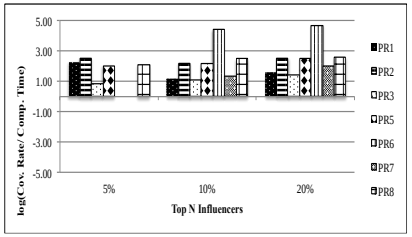
Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 2, *N*=70)



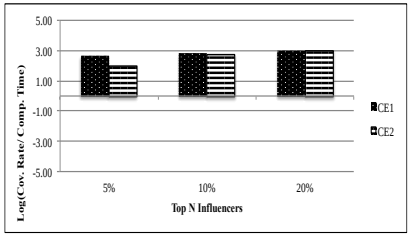
Computation Time for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



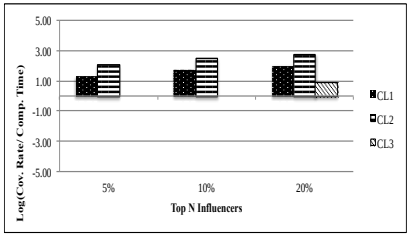
Coverage Rate/ Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



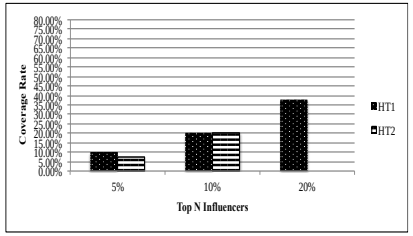
Coverage Rate/ Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



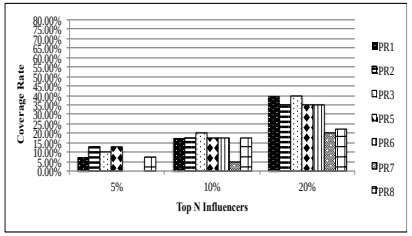
Coverage Rate/ Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 3, *N*=72)



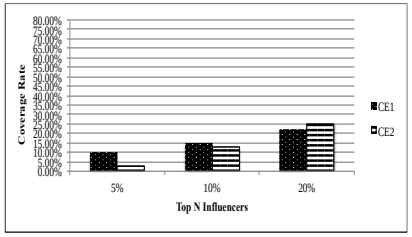
Coverage Rate/ Computation Time for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



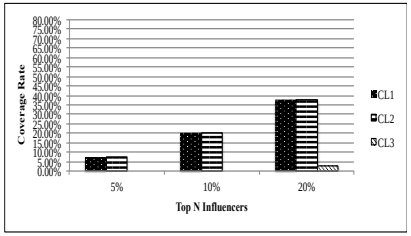
Coverage Rate for HITS-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



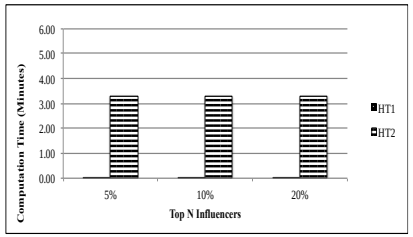
Coverage Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



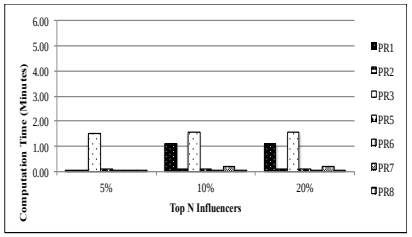
Coverage Rate for Centrality-based Methods: Twitter KY Derby Dataset (Window 3, *N*=72)



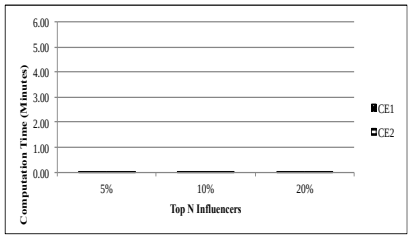
Coverage Rate for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



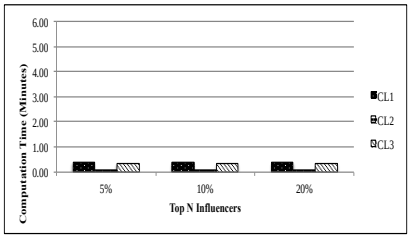
Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



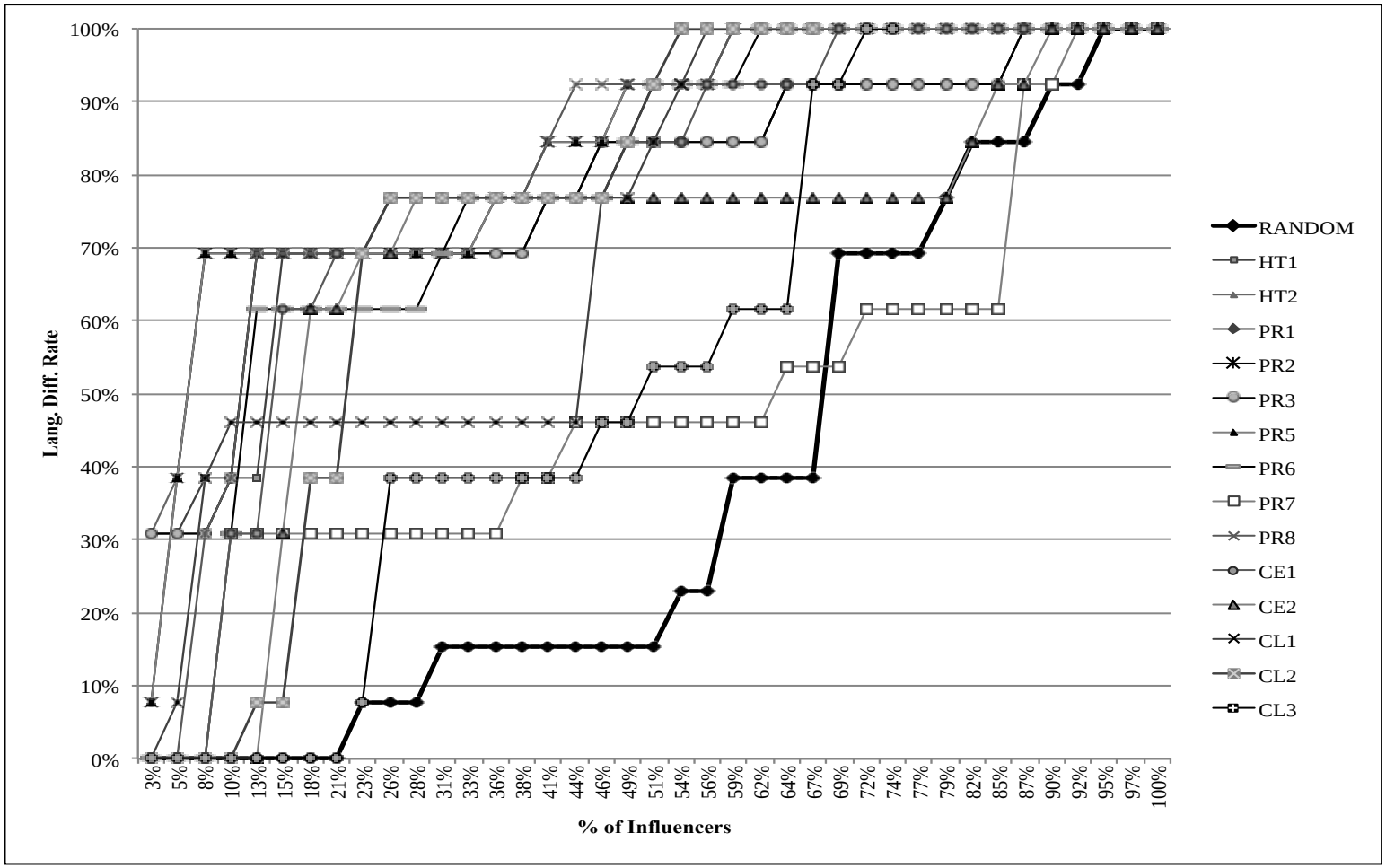
Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



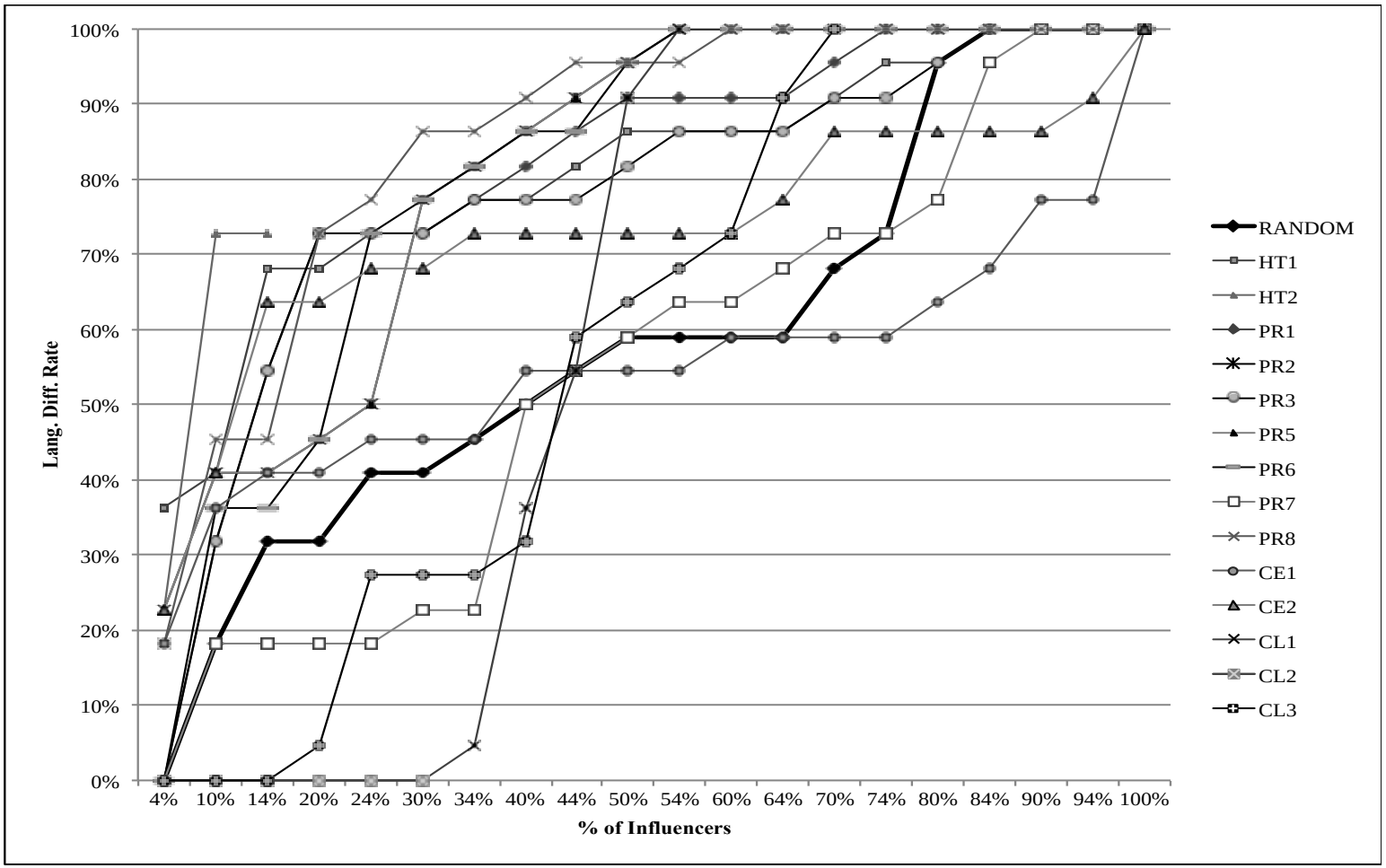
Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 3, *N*=72)



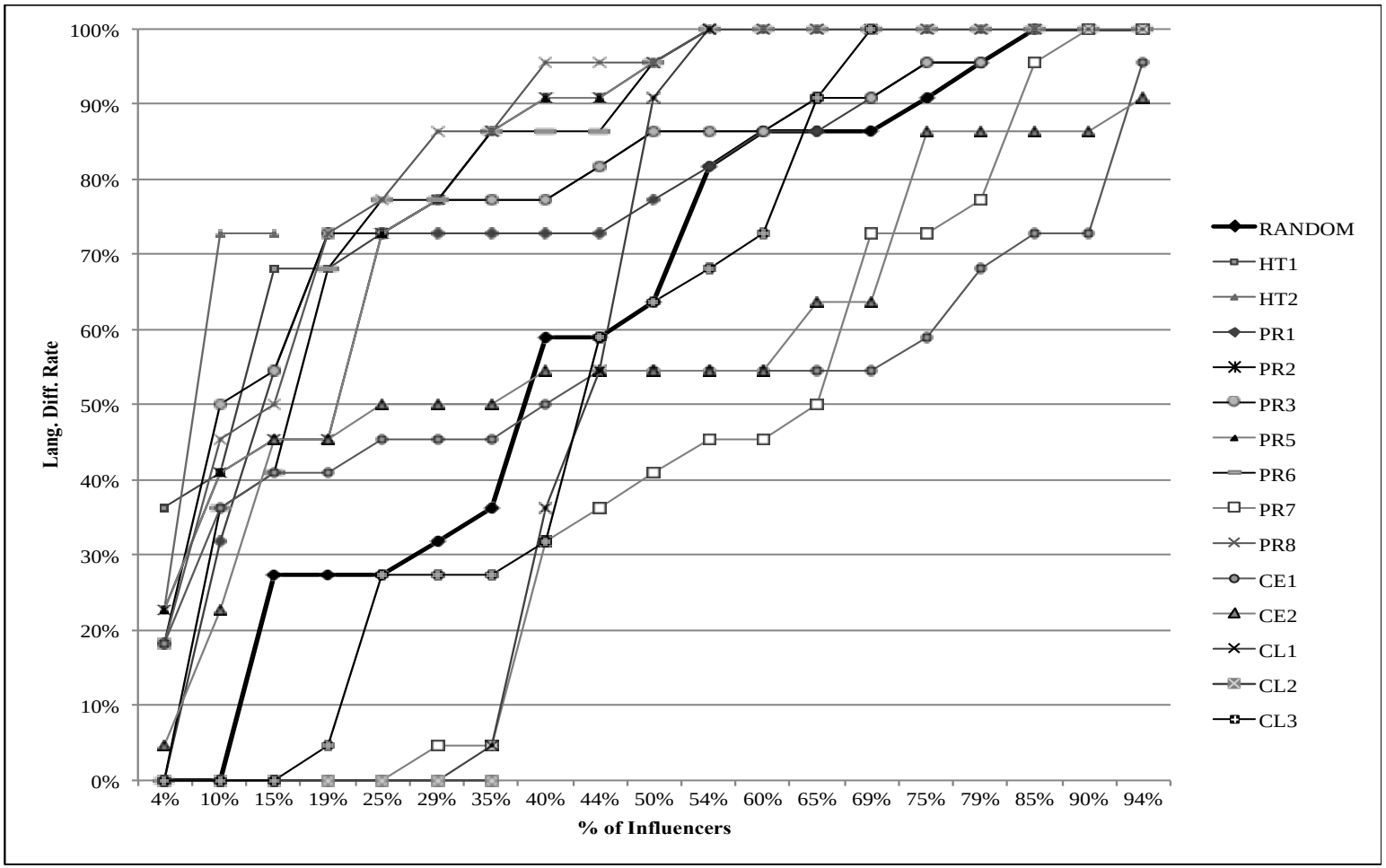
Computation Time for Clustering-based Algorithms s: Twitter KY Derby Dataset (Window 3, *N*=72)



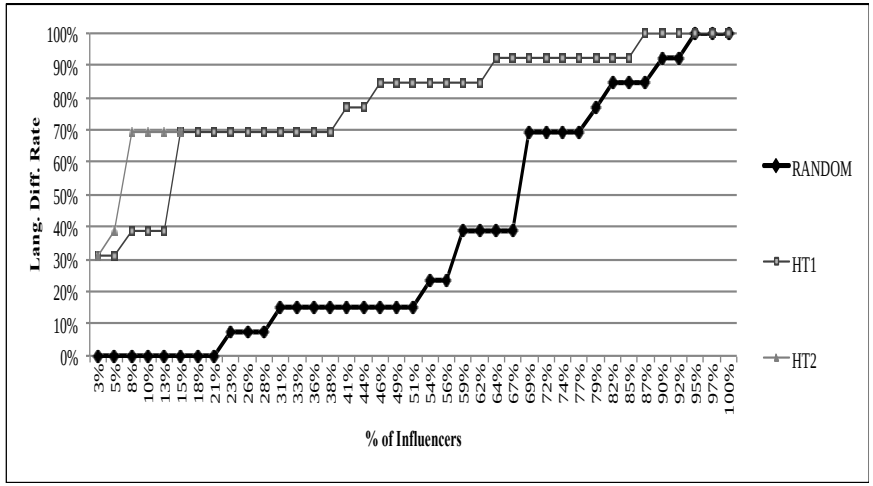
Language Diffusion Rate for Different Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



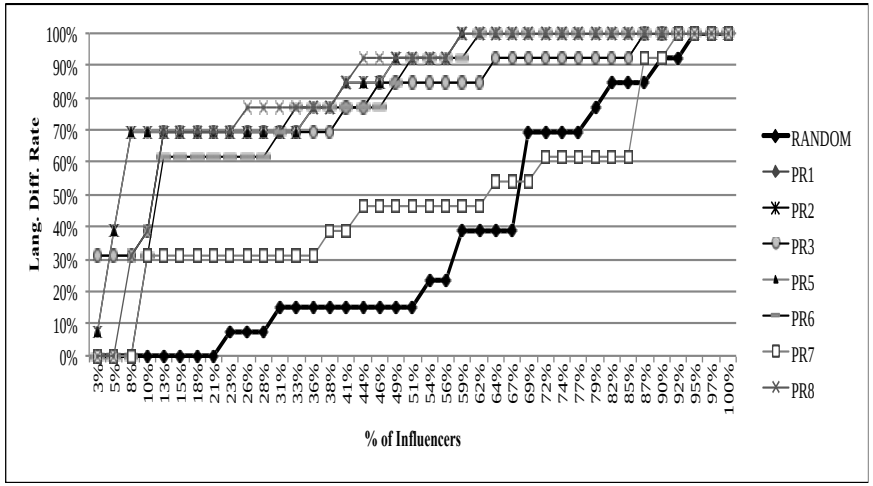
Language Diffusion Rate for Different Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



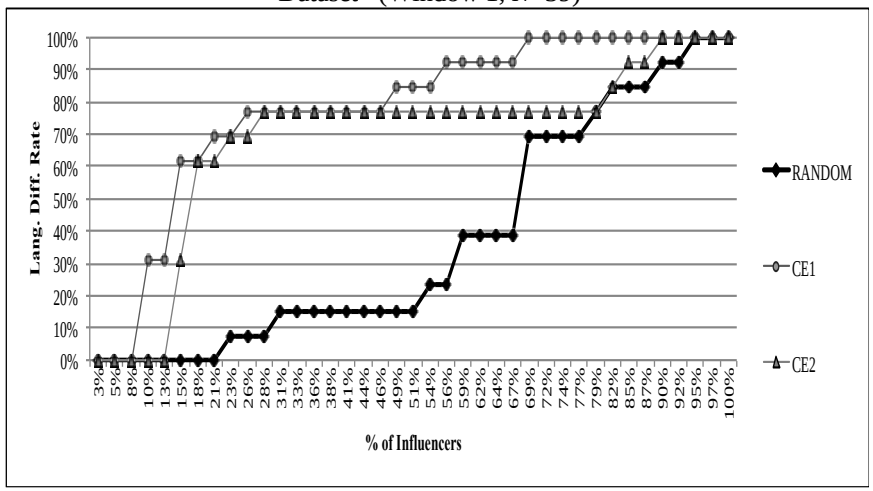
Language Diffusion Rate for Different Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



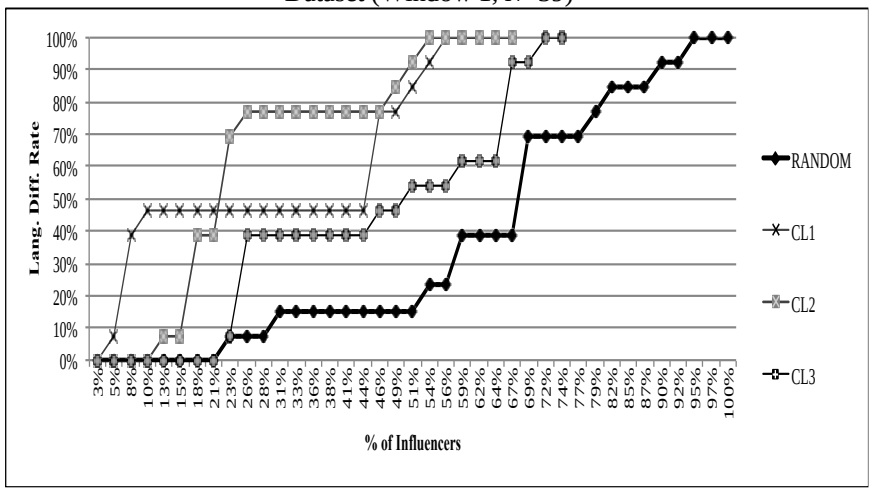
Language Diffusion Rate for HITS-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



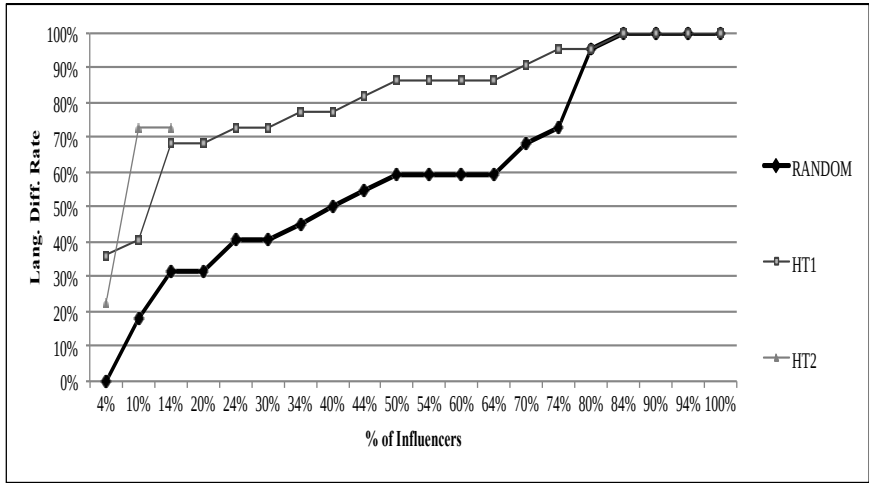
Language Diffusion Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



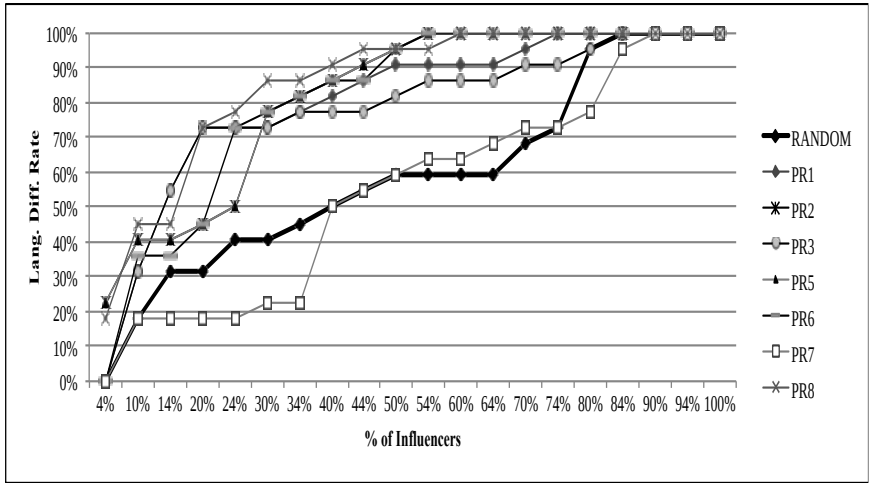
Language Diffusion Rate for Centrality_based Methods: Twitter KY Derby Dataset (Window 1, *N*=39)



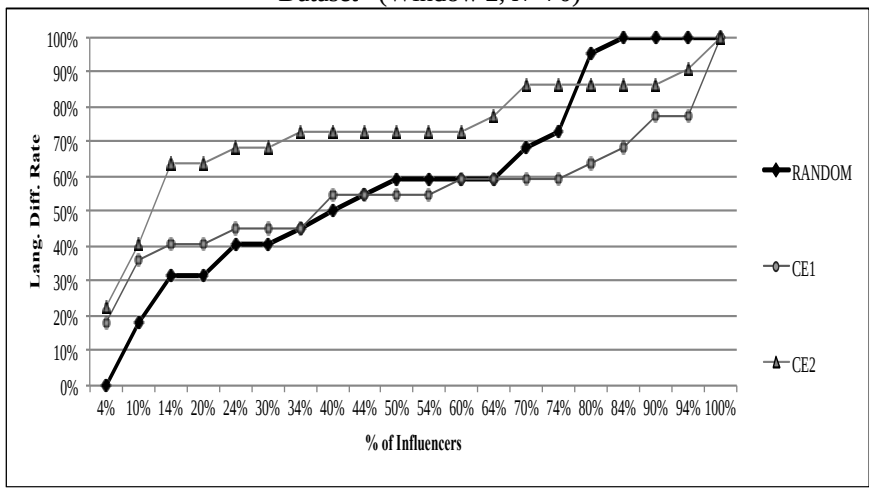
Language Diffusion Rate for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



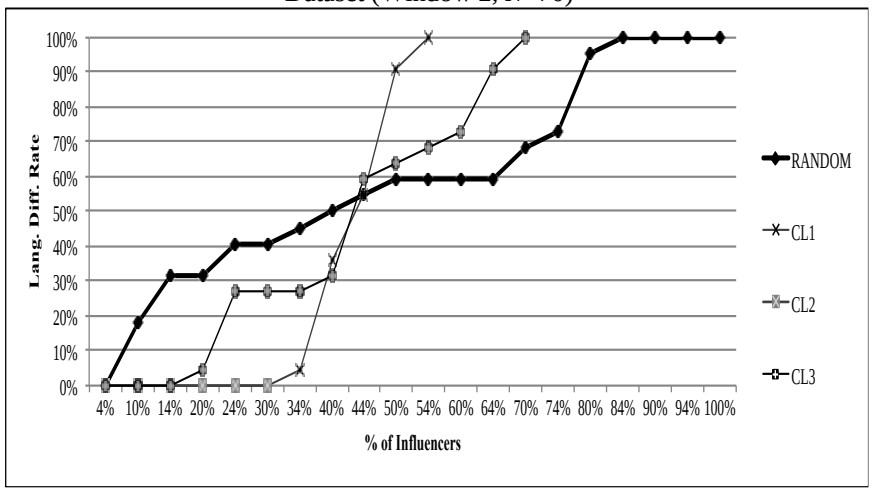
Language Diffusion Rate for HITS-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



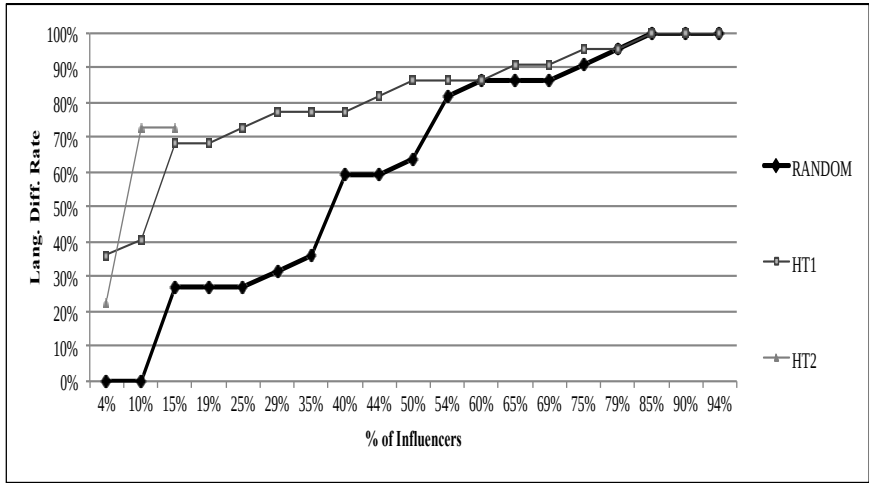
Language Diffusion Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



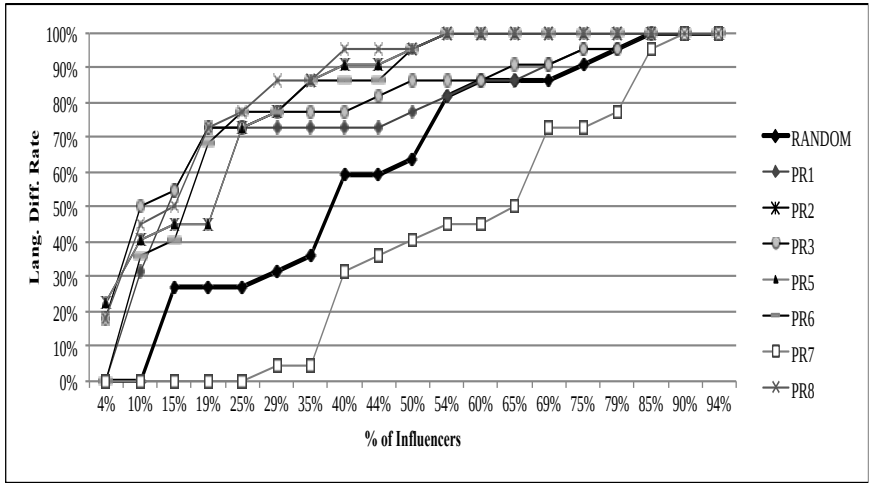
Language Diffusion Rate for Centrality_based Methods: Twitter KY Derby Dataset (Window 2, *N*=70)



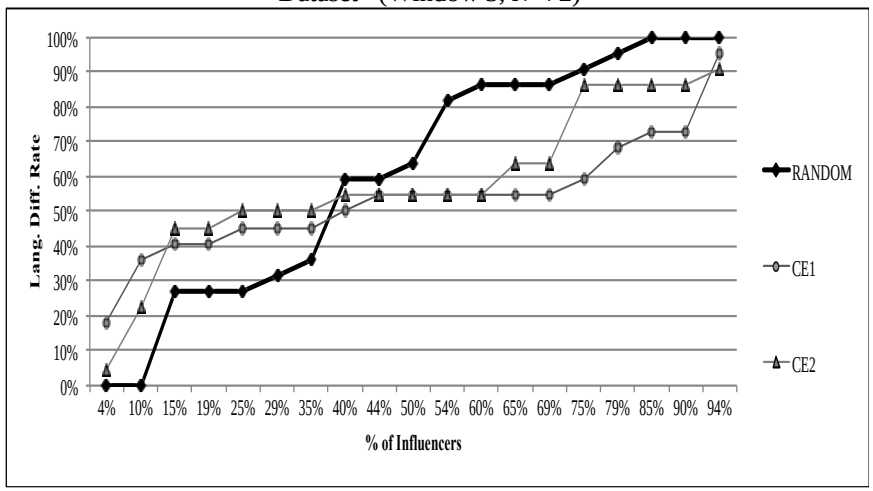
Language Diffusion Rate for Clustering -based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



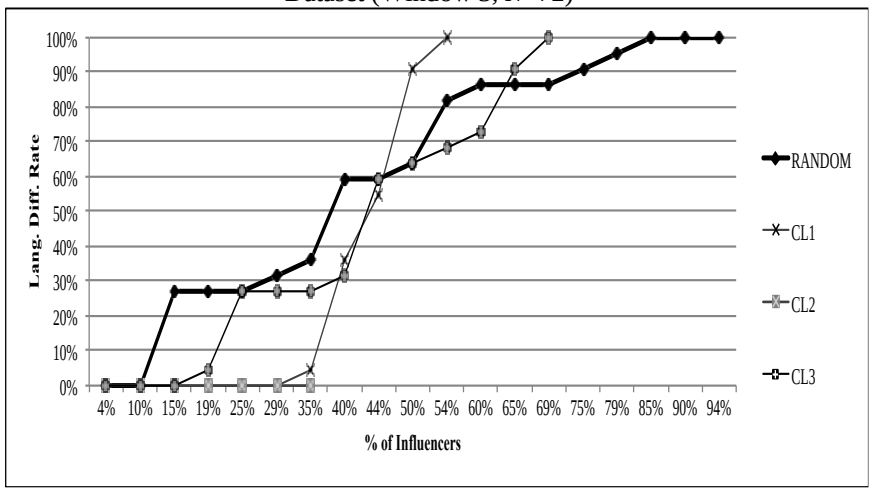
Language Diffusion Rate for HITS-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



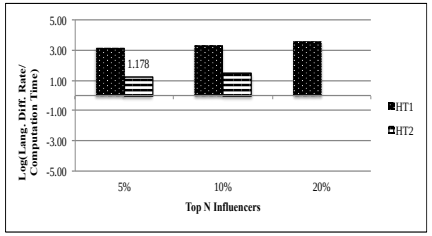
Language Diffusion Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



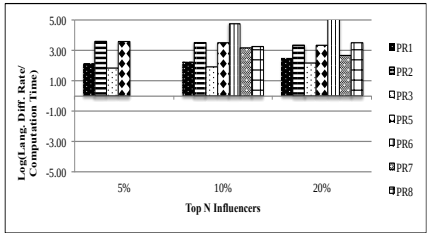
Language Diffusion Rate for Centrality_based Methods: Twitter KY Derby Dataset (Window 3, *N*=72)



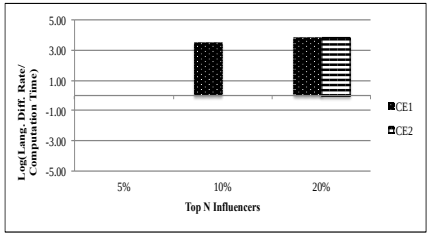
Language Diffusion Rate for Clustering -based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



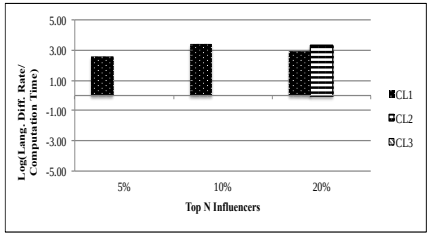
Language Diffusion Rate/ Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



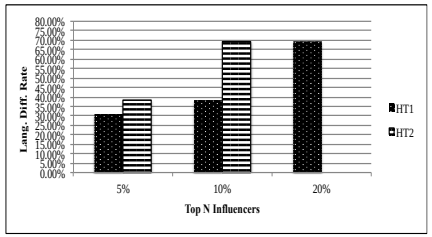
Language Diffusion Rate/ Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



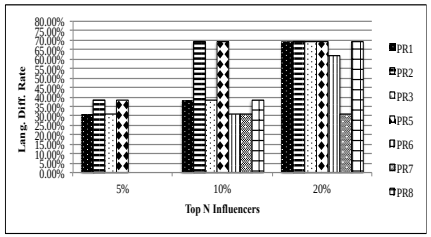
Language Diffusion Rate/ Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 1, *N*=39)



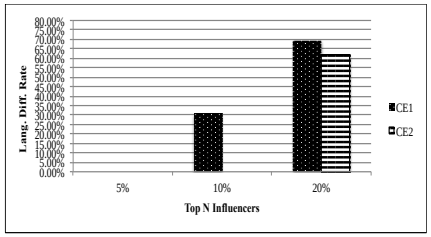
Language Diffusion Rate/ Computation Time for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



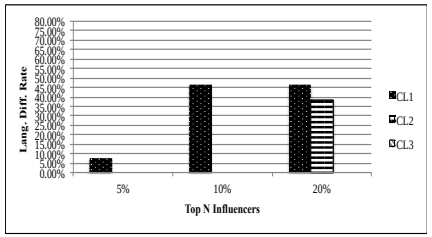
Language Diffusion Rate for HITS-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



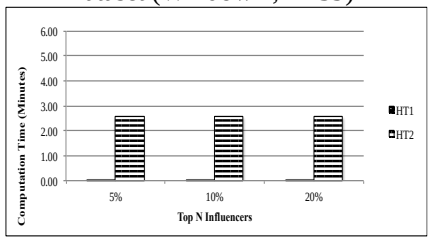
Language Diffusion Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



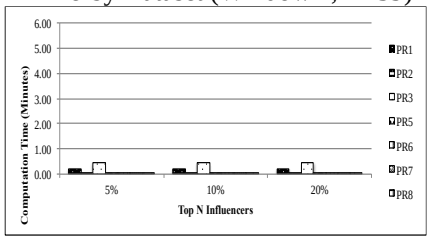
Language Diffusion Rate for Centrality-based Methods: Twitter KY Derby Dataset (Window 1, *N*=39)



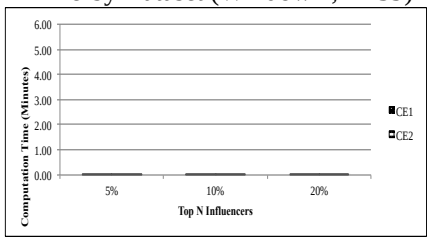
Language Diffusion Rate for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



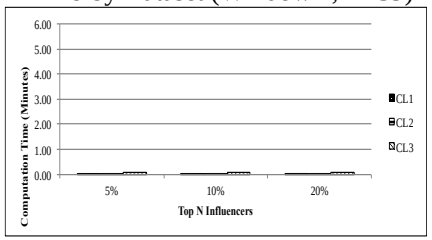
Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



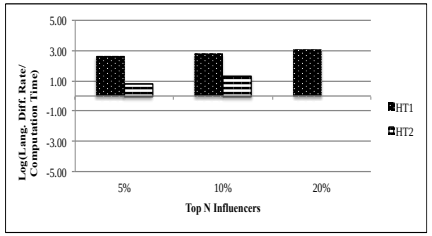
Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



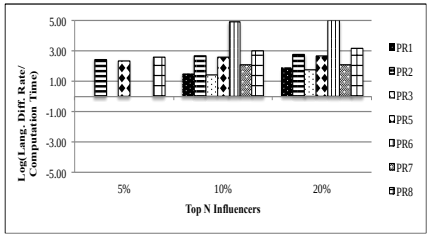
Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 1, *N*=39)



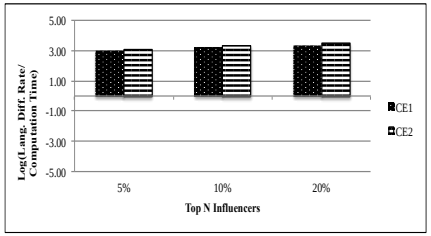
Computation Time for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



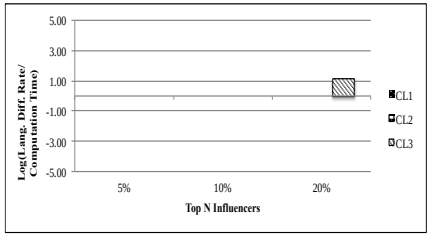
Language Diffusion Rate/ Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



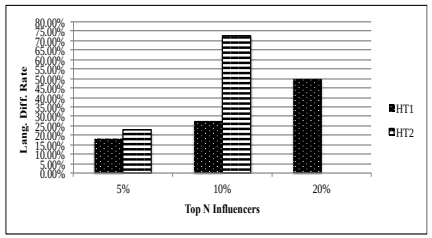
Language Diffusion Rate/ Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



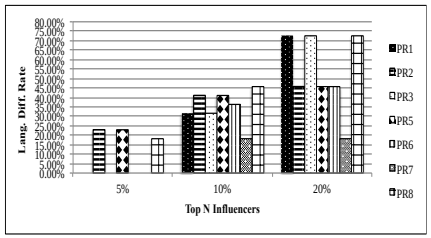
Language Diffusion Rate/ Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 2, *N*=70)



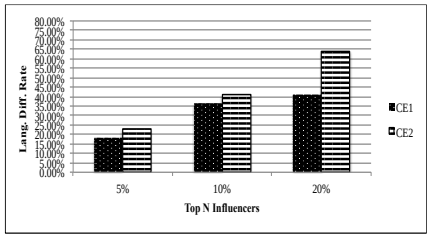
Language Diffusion Rate/ Computation Time for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



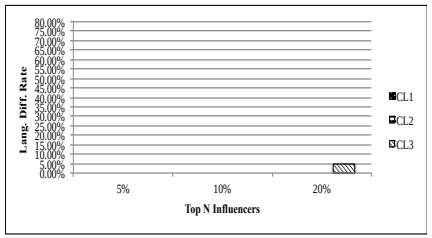
Language Diffusion Rate for HITS-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



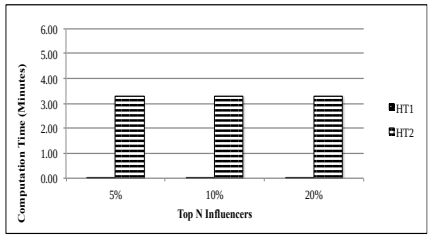
Language Diffusion Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



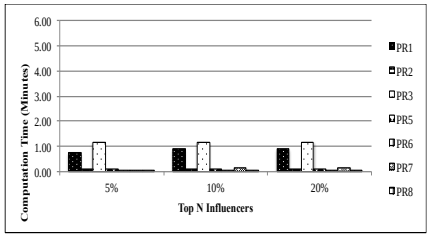
Language Diffusion Rate for Centrality-based Methods: Twitter KY Derby Dataset (Window 2, *N*=70)



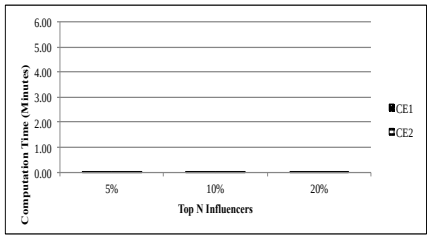
Language Diffusion Rate for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



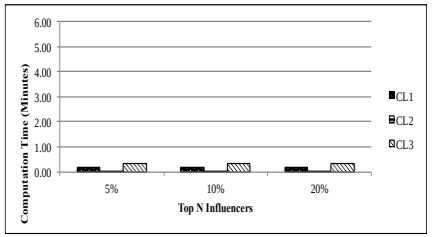
Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



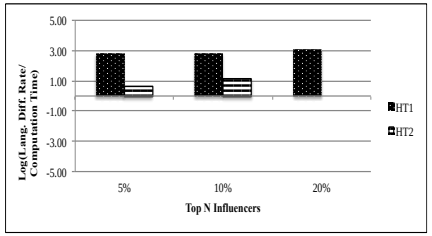
Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



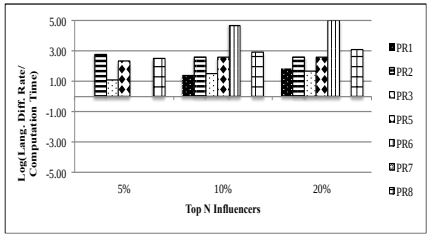
Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 2, *N*=70)



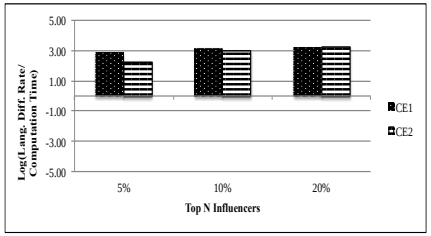
Computation Time for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



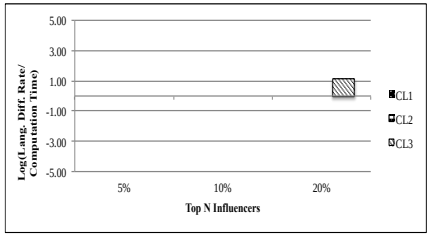
Language Diffusion Rate/ Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



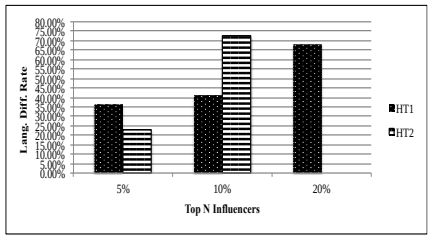
Language Diffusion Rate/ Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



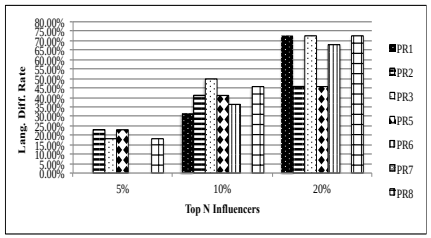
Language Diffusion Rate/ Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 3, *N*=72)



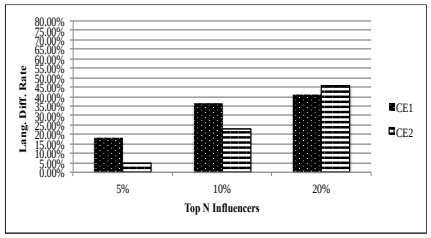
Language Diffusion Rate/ Computation Time for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



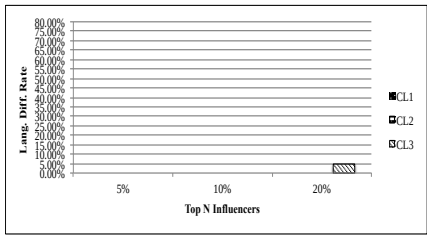
Language Diffusion Rate for HITS-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



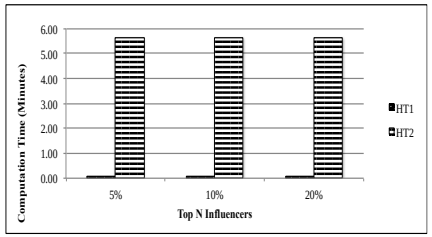
Language Diffusion Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



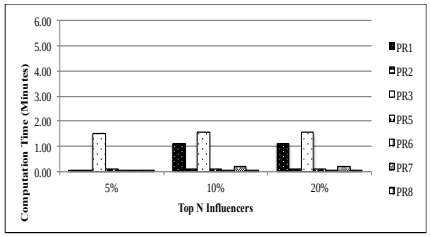
Language Diffusion Rate for Centrality-based Methods: Twitter KY Derby Dataset (Window 3, *N*=72)



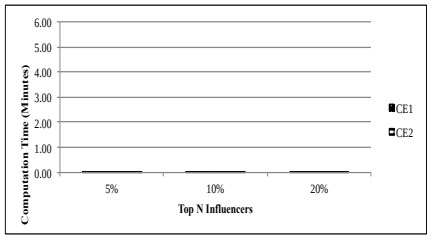
Language Diffusion Rate for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



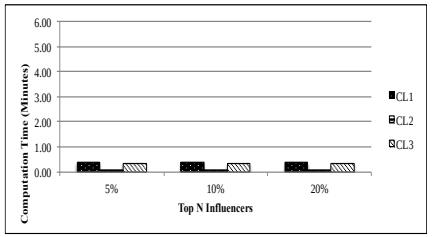
Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



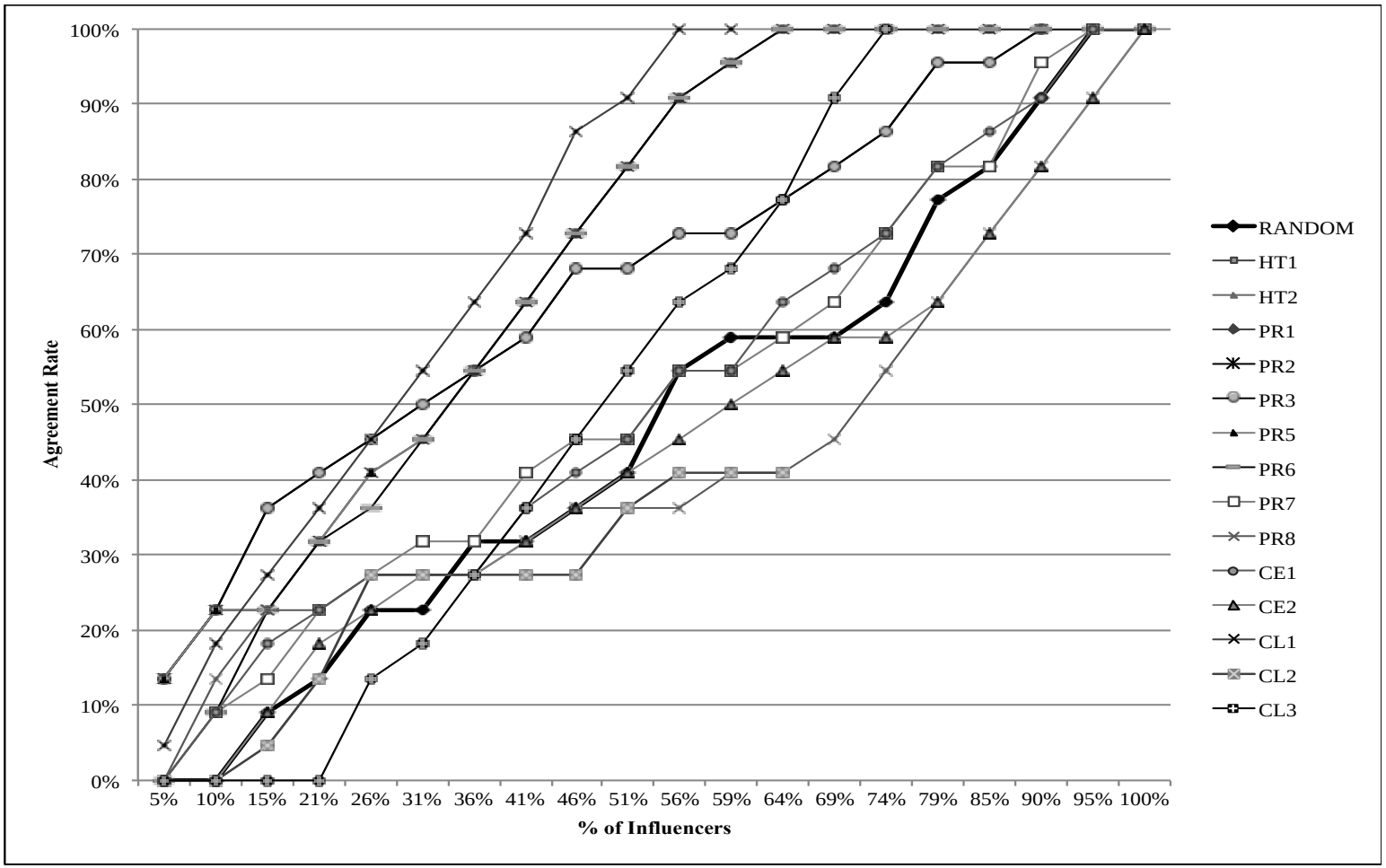
Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



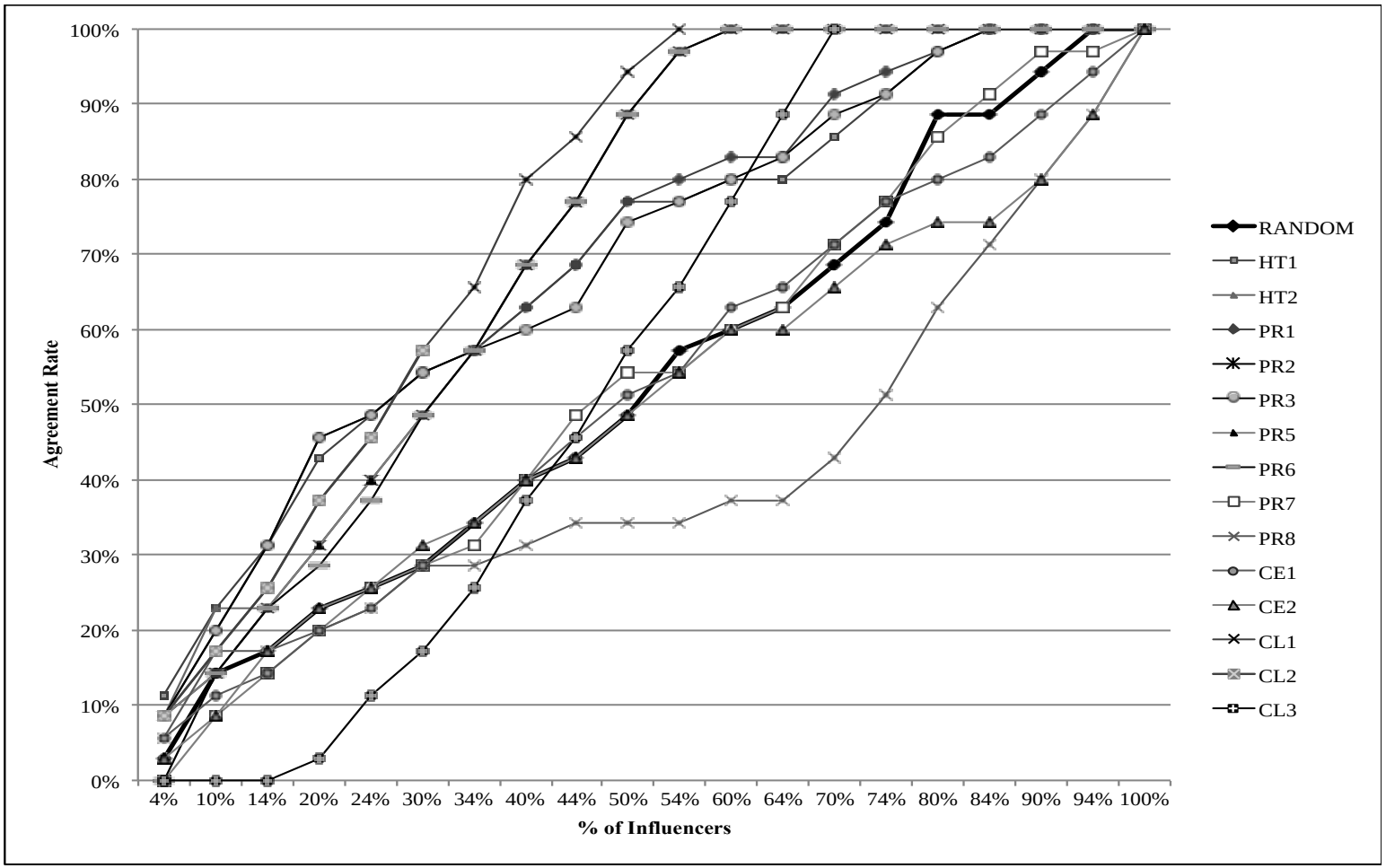
Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 3, *N*=72)



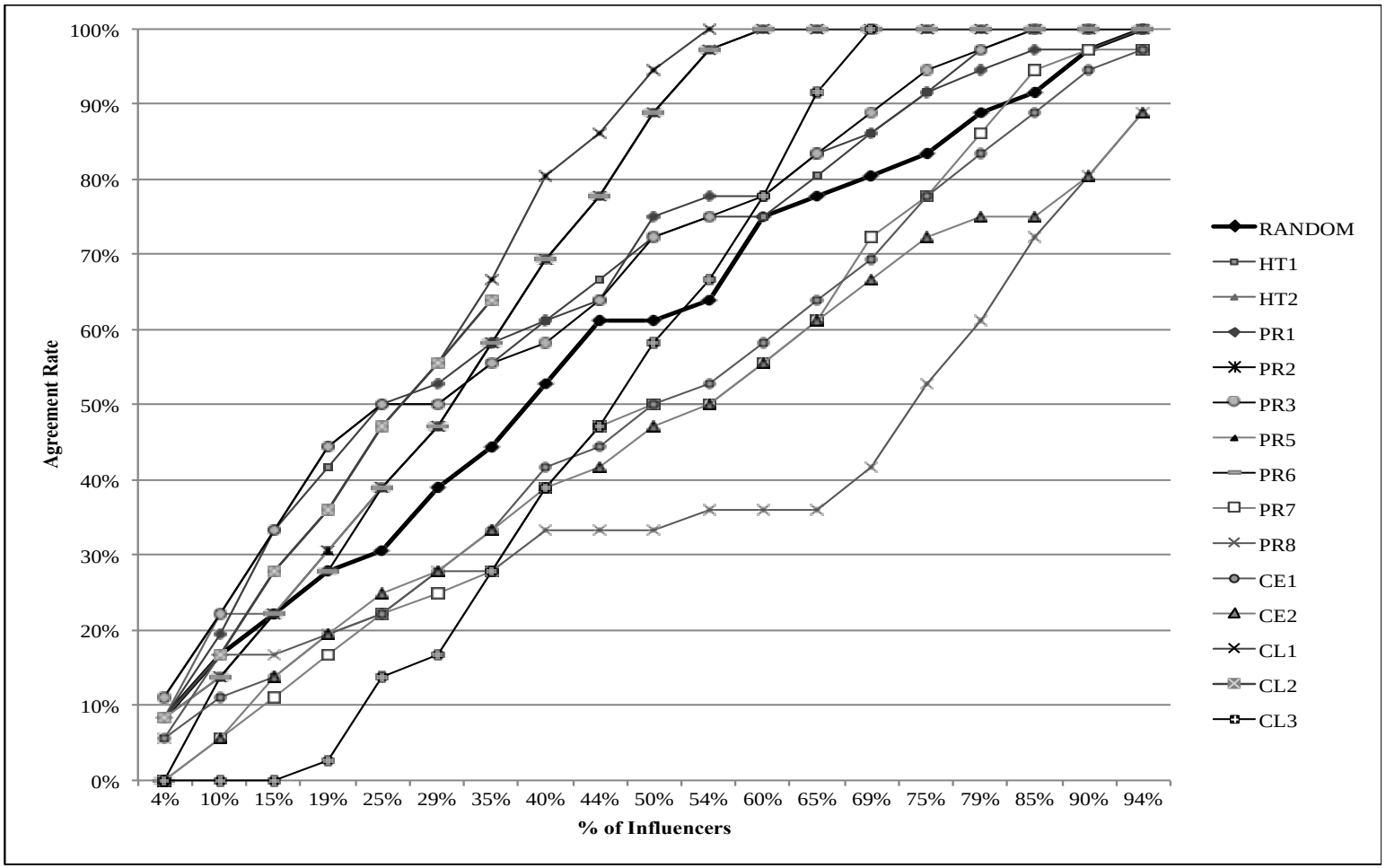
Computation Time for Clustering-based Algorithms s: Twitter KY Derby Dataset (Window 3, *N*=72)



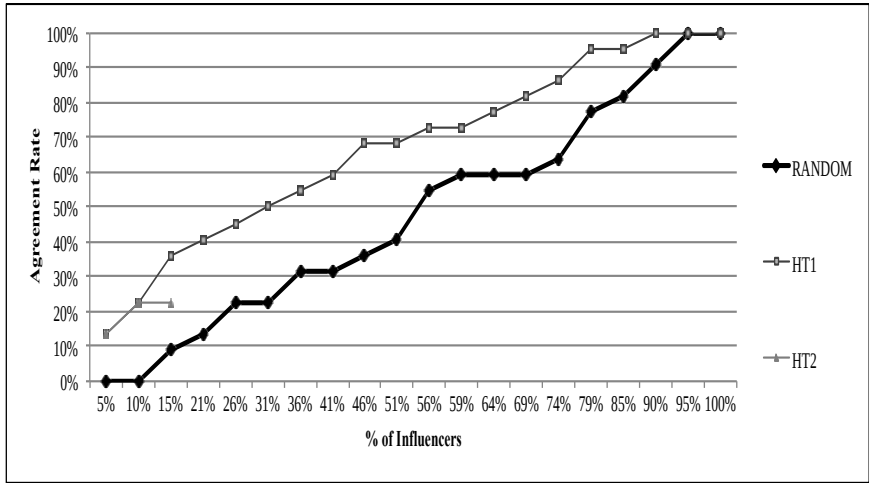
Agreement Rate for Different Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



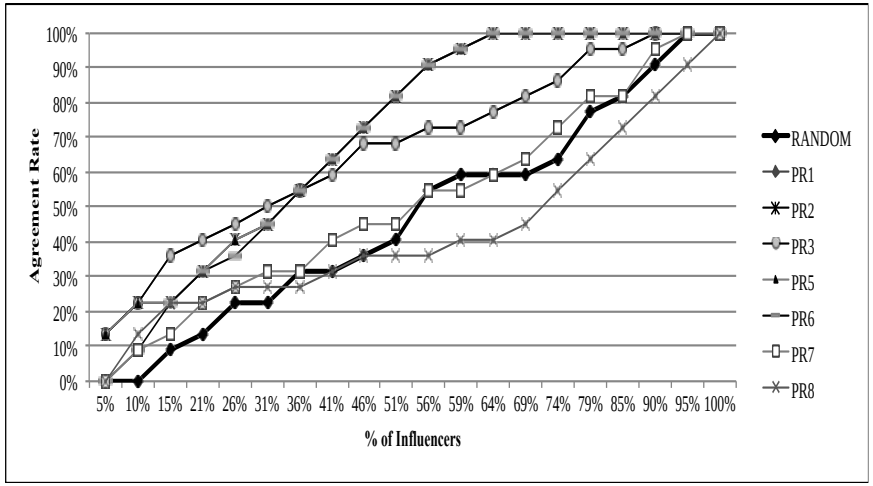
Agreement Rate for Different Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



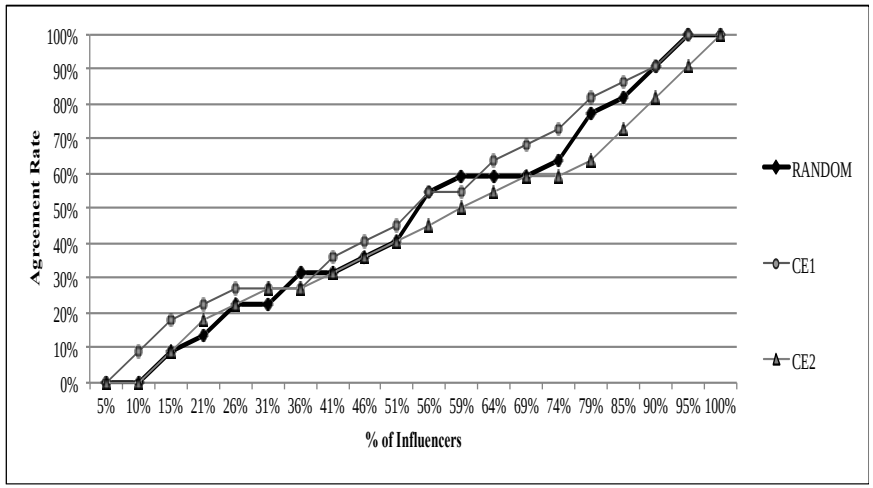
Agreement Rate for Different Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



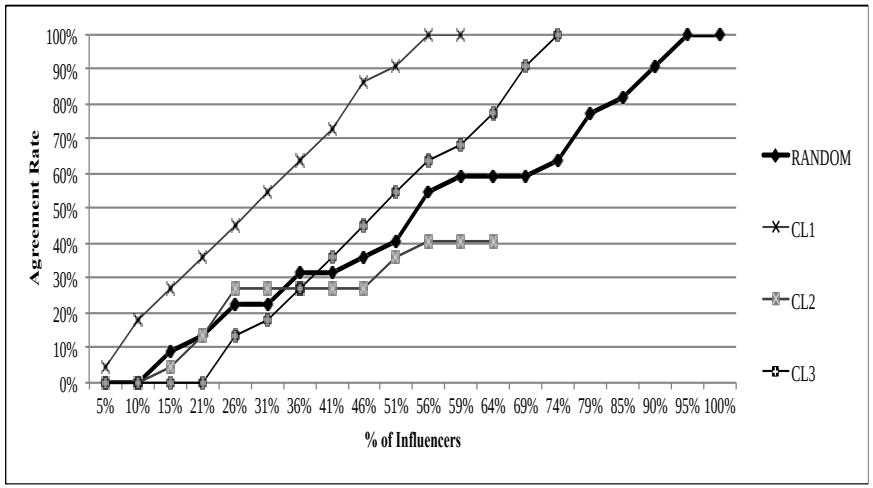
Agreement Rate for HITS-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



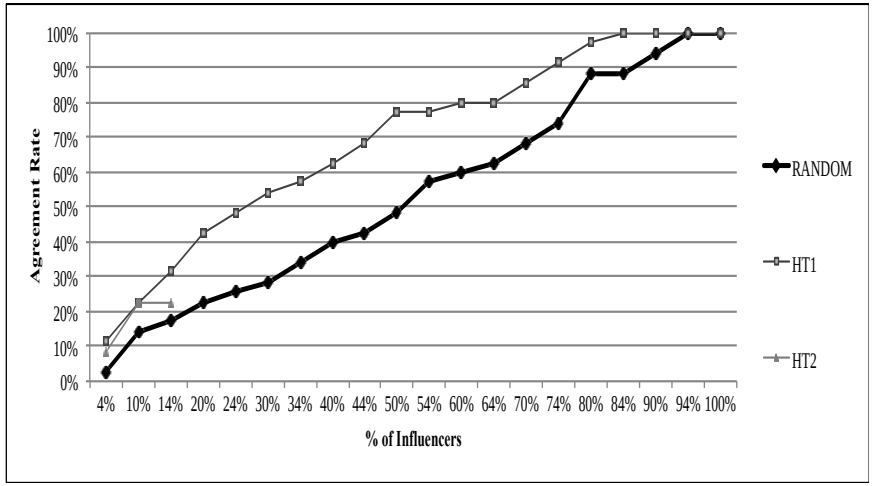
Agreement Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



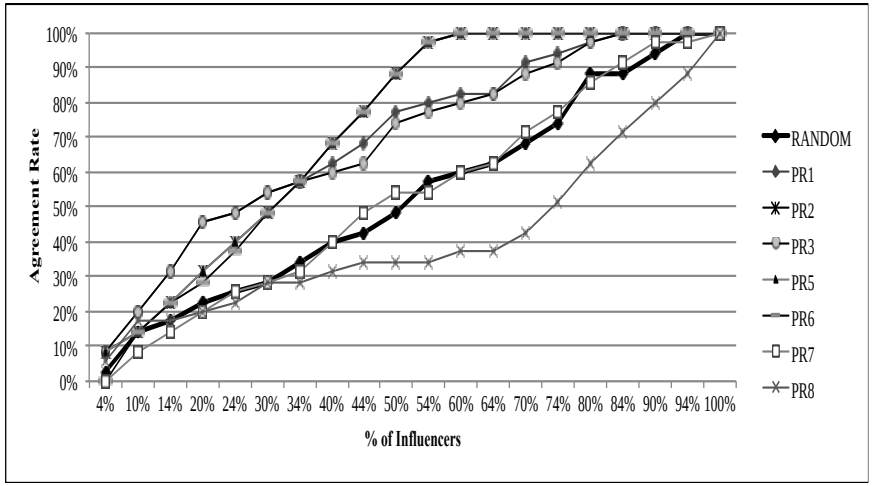
Agreement Rate for Centrality_based Methods: Twitter KY Derby Dataset (Window 1, *N*=39)



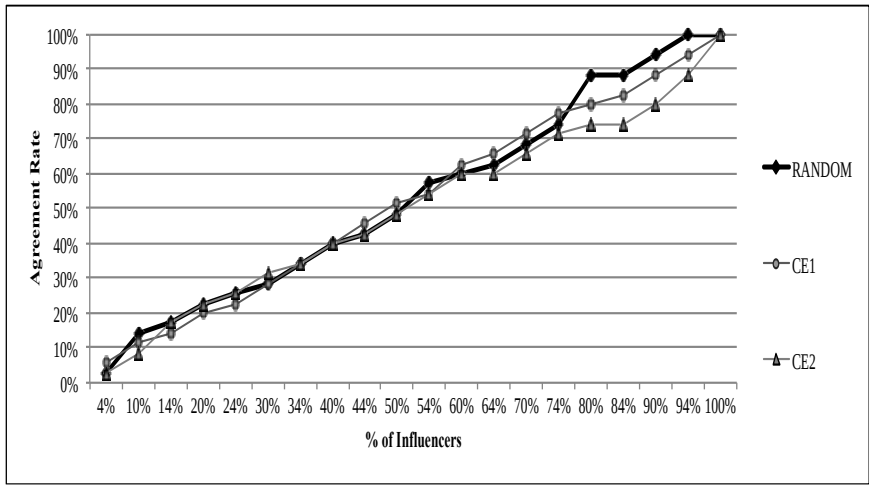
Agreement Rate for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



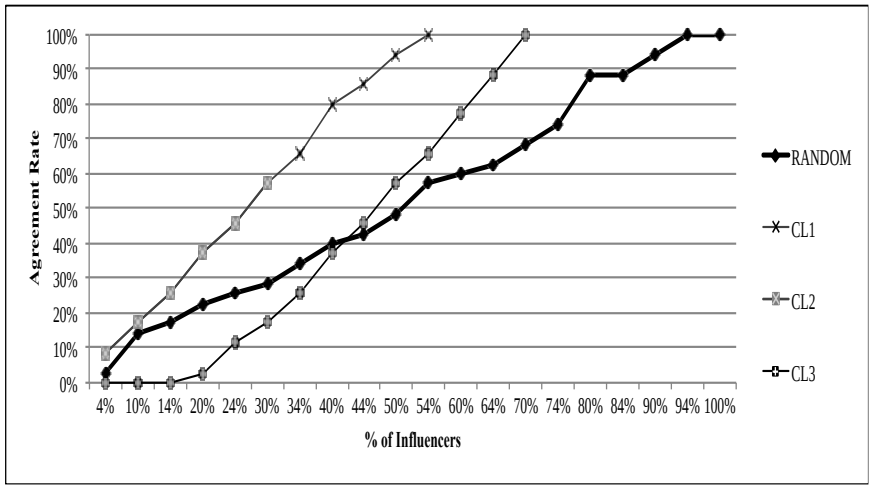
Agreement Rate for HITS-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



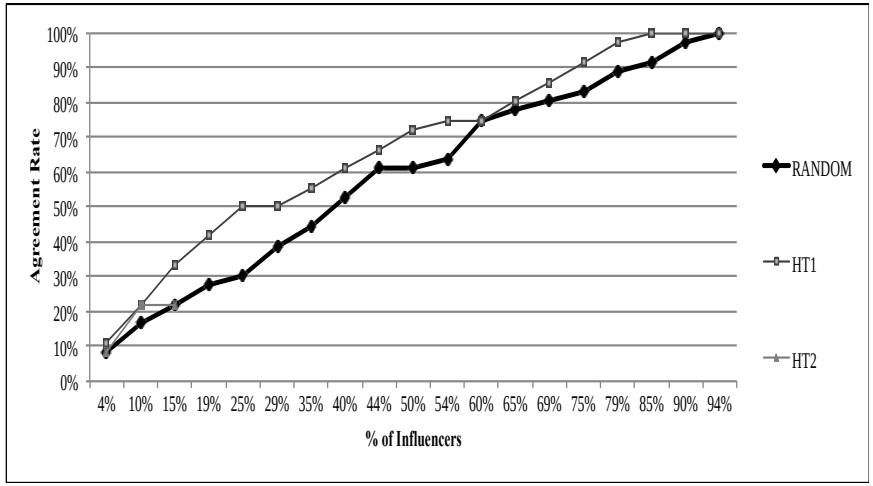
Agreement Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



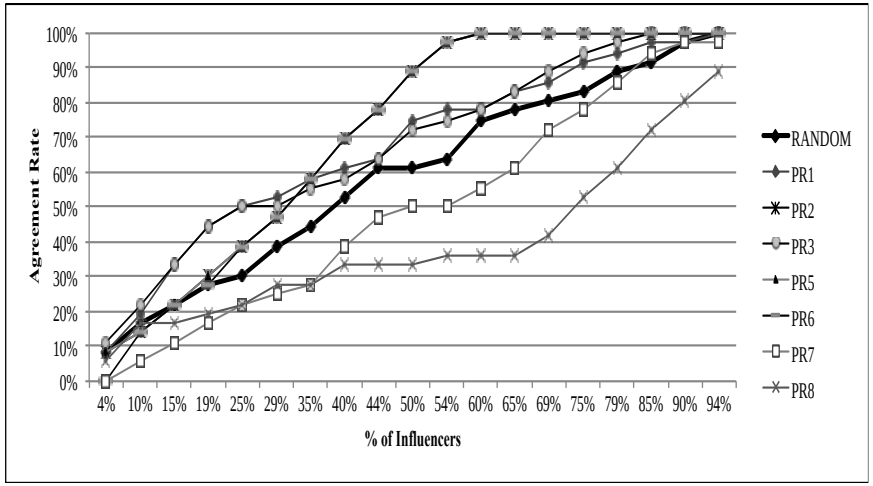
Agreement Rate for Centrality_based Methods: Twitter KY Derby Dataset (Window 2, *N*=70)



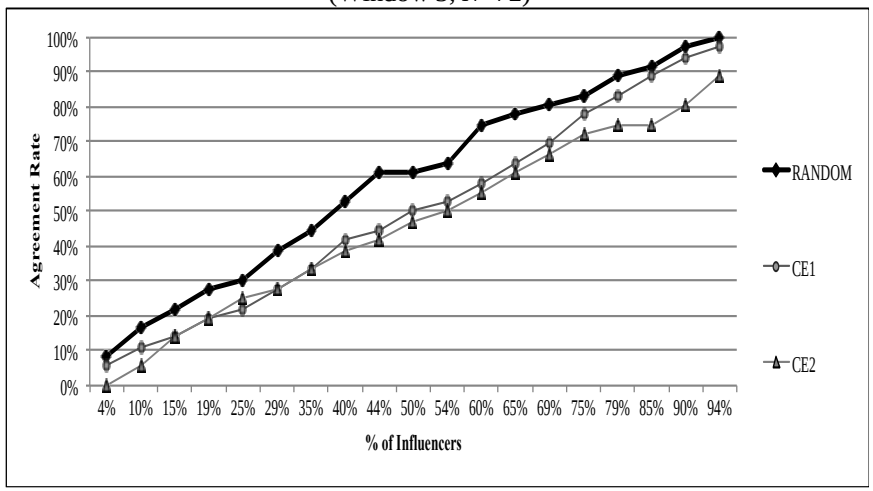
Agreement Rate for Clustering -based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



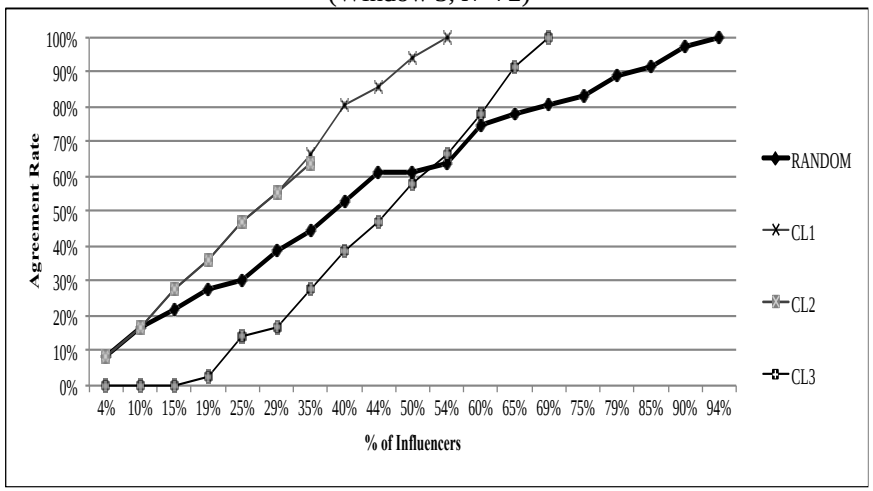
Agreement Rate for HITS-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



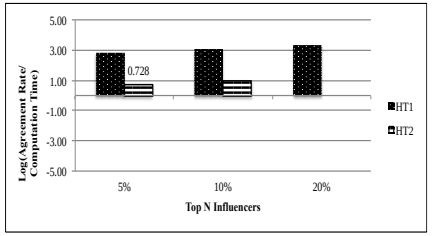
Agreement Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



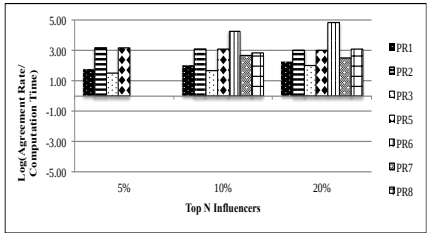
Agreement Rate for Centrality_based Methods: Twitter KY Derby Dataset (Window 3, *N*=72)



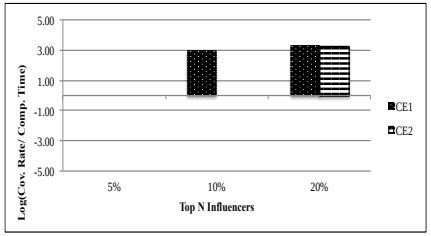
Agreement Rate for Clustering -based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



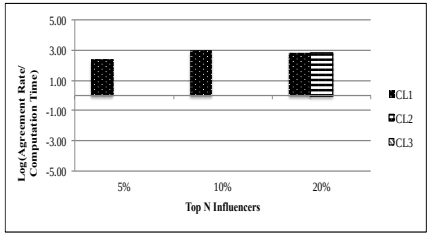
Agreement Rate/ Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



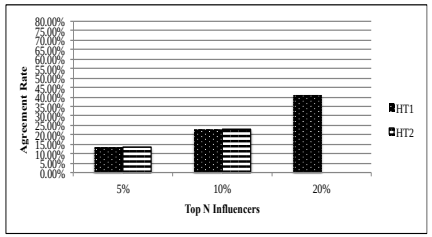
Agreement Rate/ Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



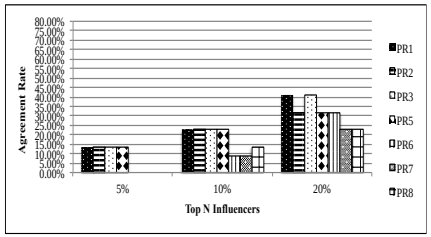
Agreement Rate/ Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 1, *N*=39)



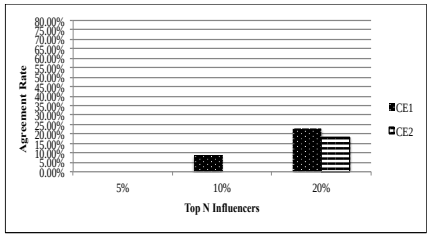
Agreement Rate/ Computation Time for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



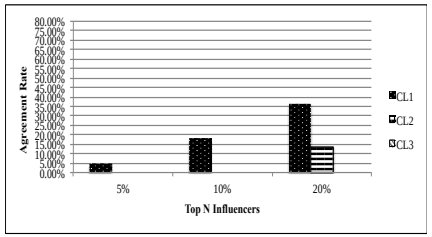
Agreement Rate for HITS-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



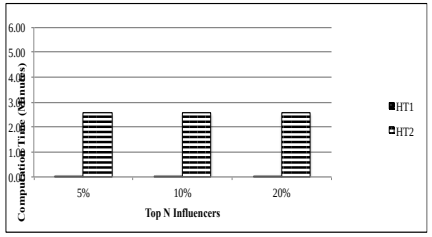
Agreement Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



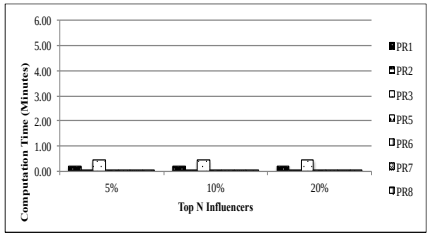
Agreement Rate for Centrality-based Methods: Twitter KY Derby Dataset (Window 1, *N*=39)



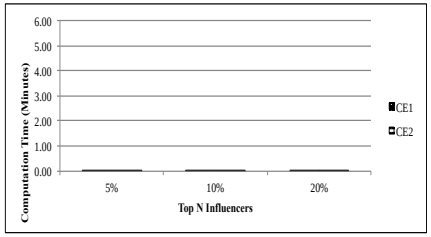
Agreement Rate for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



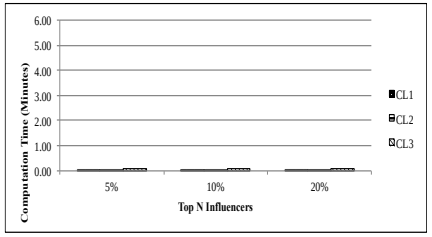
Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



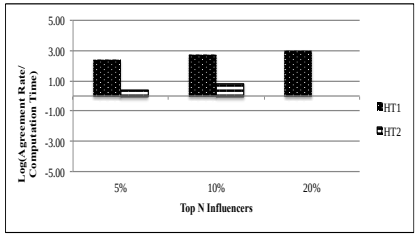
Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



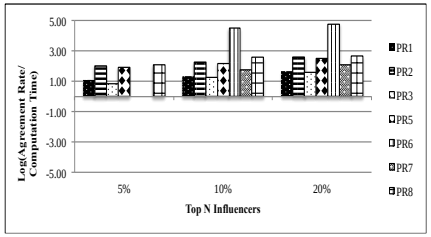
Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 1, *N*=39)



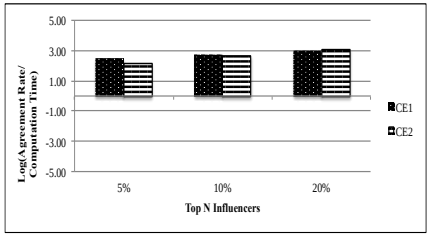
Computation Time for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 1, *N*=39)



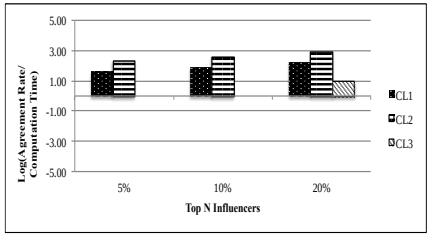
Agreement Rate/ Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



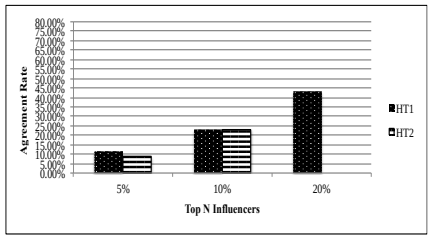
Agreement Rate/ Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



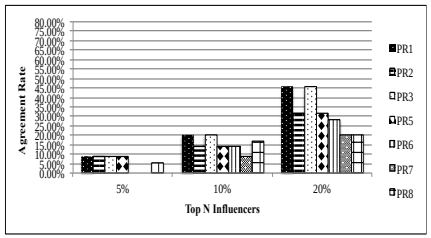
Agreement Rate/ Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 2, *N*=70)



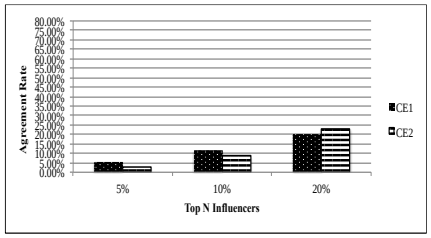
Agreement Rate/ Computation Time for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



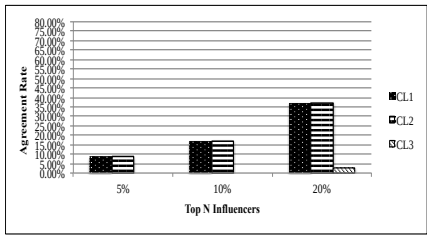
Agreement Rate for HITS-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



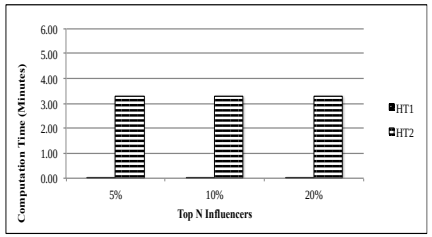
Agreement Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



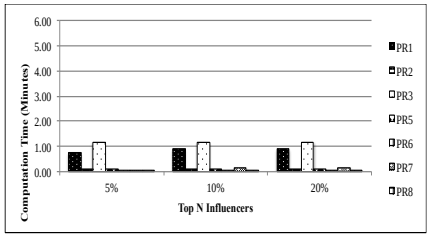
Agreement Rate for Centrality-based Methods: Twitter KY Derby Dataset (Window 2, *N*=70)



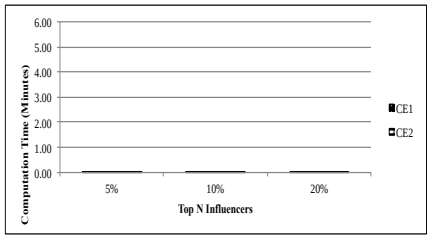
Agreement Rate for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



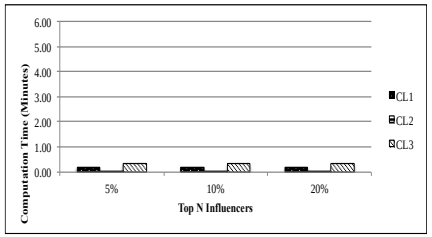
Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



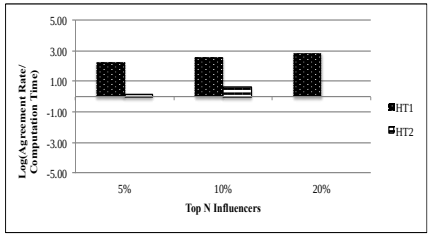
Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



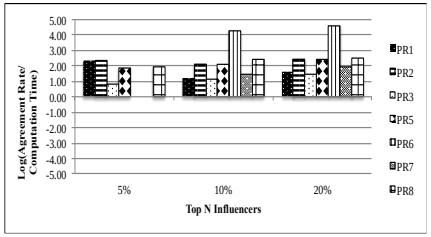
Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 2, *N*=70)



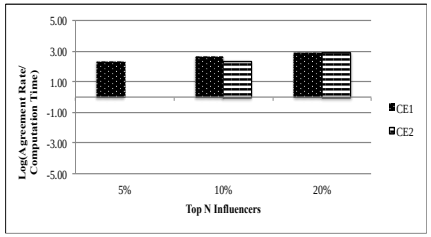
Computation Time for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 2, *N*=70)



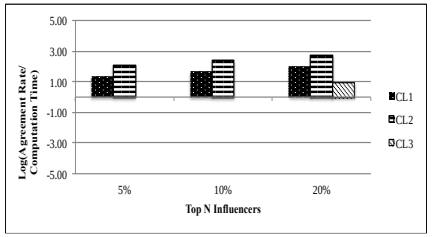
Agreement Rate/ Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



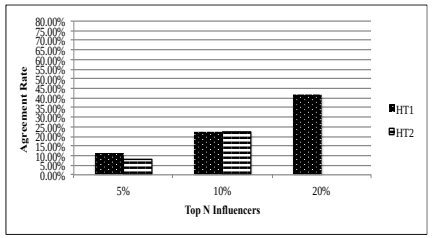
Agreement Rate/ Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



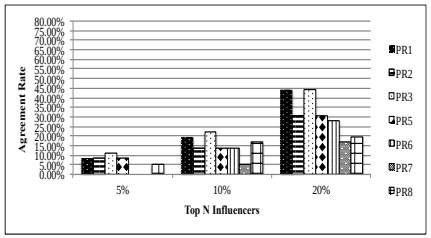
Agreement Rate/ Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 3, *N*=72)



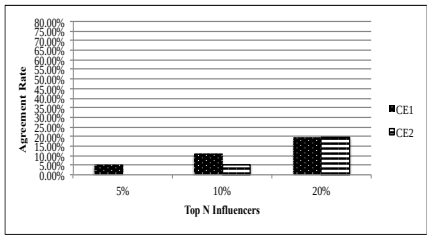
Agreement Rate/ Computation Time for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



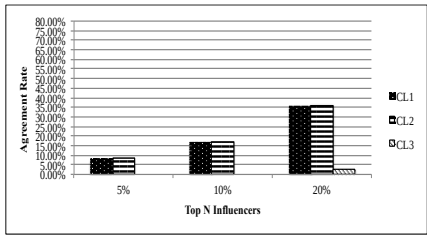
Agreement Rate for HITS-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



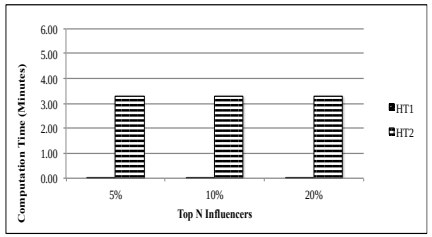
Agreement Rate for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



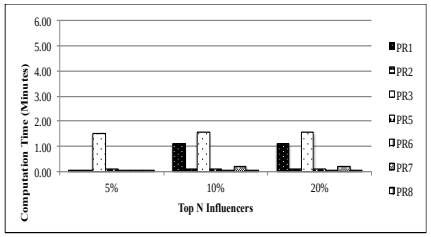
Agreement Rate for Centrality-based Methods: Twitter KY Derby Dataset (Window 3, *N*=72)



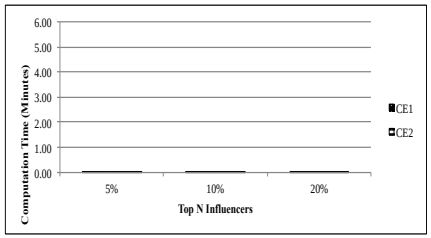
Agreement Rate for Clustering-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



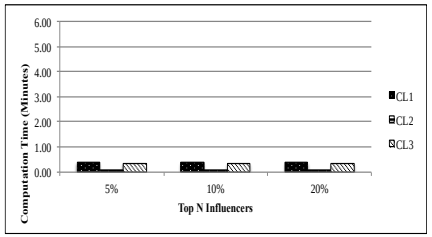
Computation Time for HITS-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



Computation Time for PageRank-based Algorithms: Twitter KY Derby Dataset (Window 3, *N*=72)



Computation Time for Centrality-based Methods: Twitter KY Derby Dataset (Window 3, *N*=72)



Computation Time for Clustering-based Algorithms s: Twitter KY Derby Dataset (Window 3, *N*=72)

References

Abrahams, A. S., Jiao, J., Fan, W., Wang, G. A., & Zhang, Z. (2013). What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings. *Decision Support Systems*, *55*(4), 871–882.

Ahituv, N., Zif, J., & Machlin, I. (1998). Environmental scanning and information systems in relation to success in introducing new products. *Information & Management*, *33*(4), 201–211.

Amblee, N., & Bui, T. (2011). Harnessing the influence of social proof in online shopping: The effect of electronic word of mouth on sales of digital microproducts. *International Journal of Electronic Commerce*, *16*(2), 91–114.

APCO Worldwide and Gagen McDonald, The 3rd Annual Employee Engagement Survey. (2011).

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (Vol. 1, pp. 492–499). Toronto, Canada: IEEE.

Barbieri, D., Braga, D., Ceri, S., Valle, E. Della, Huang, Y., Tresp, V., … Wermser, H. (2010). Deductive and inductive stream reasoning for semantic social media analytics. *Intelligent Systems, IEEE*, *25*(6), 32–41.

Barnes, L. L. (2011). Social Bookmarking Sites: A Review. *Collaborative Librarianship*.

Best, D. M., Bruce, J., Dowson, S., Love, O., & McGrath, L. (2012). Web-based visual analytics for social media. In *AAAI ICWSM SocMedVis: Workshop on Social Media Visualization (AAAI Technical Report WS-12-03)* (pp. 2–5). Dublin, Ireland.

Bickart, B., & Schindler, R. M. (2001). Internet forums as influential sources of consumer information. *Journal of Interactive Marketing*, *15*(3), 31–40.

Black, J. S. (1982). Opinion leaders: Is anyone following? *Public Opinion Quarterly*, *46*(2), 169–176.

Blood, R. (2002). *The weblog handbook: Practical advice on creating and maintaining your blog*. Basic Books.

Boden, C., Karnstedt, M., Fernandez, M., & Markl, V. (2013). Large-scale social-media analytics on stratosphere. In *Proceedings of the 2013 22nd international conference on World Wide Web companion* (pp. 257–260). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, *27*(1), 55–71.

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2013). *Analyzing social networks*. SAGE Publications Limited.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, *323*(5916), 892–895.

Brown, A. D., & Humphreys, M. (2003). Epic and tragic tales Making sense of change. *The Journal of Applied Behavioral Science*, *39*(2), 121–144.

Butler, B. S. (2001). Membership size, communication activity, and sustainability: A resource-based model of online social structures. *Information Systems Research*, *12*(4), 346–362.

Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D. S., & Ertl, T. (2012). Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 143–152). Seattle, WA, USA: IEEE.

Chan, K. K., & Misra, S. (1990). Characteristics of the opinion leader: A new dimension. *Journal of Advertising*, *19*(3), 53–60.

Chau, M., & Xu, J. (2012). Business intelligence in blogs: Understanding consumer interactions and communities. *MIS Quarterly*, *36*(4), 1189–1216.

Chen, W., Cheng, S., He, X., & Jiang, F. (2012). Influencerank: An efficient social influence measurement for millions of users in microblog. In *Cloud and Green Computing (CGC), 2012 Second International Conference on* (pp. 563–570). IEEE.

Cheng, F.-T., Yan, C., Huang, Y.-P., & Zhou, L. (2012). Algorithm of identifying opinion leaders in BBS. In *Cloud Computing and Intelligent Systems (CCIS), 2012 IEEE 2nd International Conference on* (Vol. 3, pp. 1149–1152). IEEE.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*(3), 345–354.

Colbaugh, R., & Glass, K. (2011). Detecting emerging topics and trends via social media analytics. In *Proceedings of the 2011 IADIS International Conference e-Commerce* (p. 51). Rome, Italy.

Cooper, T. (2006). Enhancing insight discovery by balancing the focus of analytics between strategic and tactical levels. *The Journal of Database Marketing & Customer Strategy Management*, *13*(4), 261–270.

Daft, R. L., Sormunen, J., & Parks, D. (1988). Chief executive scanning, environmental characteristics, and company performance: An empirical study. *Strategic Management Journal*, *9*(2), 123–139.

Dean, J., & Ghemawat, S. (2010). MapReduce: a flexible data processing tool. *Communications of the ACM*, *53*(1), 72–77.

Dellarocas, C., Gao, G., & Narayan, R. (2010). Are consumers more likely to contribute online reviews for hit or niche products? *Journal of Management Information Systems*, *27*(2), 127–158.

Deloitte. (2012). *Global Risk Management Survey, Eighth Edition: Setting a Higher Bar*. Retrieved from http://www.deloitte.com/assets/Dcom-UnitedStates/Local Assets/Documents/AERS/us_aers_grr_grms8_infographic_pdf_072313.pdf

Dervin, B. (1998). Sense-making theory and practice: an overview of user interests in knowledge seeking and use. *Journal of Knowledge Management*, *2*(2), 36–46.

Dill, D. D., & Friedman, C. P. (1979). An analysis of frameworks for research on innovation and change in higher education. *Review of Educational Research*, *49*(3), 411–435.

Dittrich, J., & Quiané-Ruiz, J.-A. (2012). Efficient big data processing in Hadoop MapReduce. *Proceedings of the VLDB Endowment*, *5*(12), 2014–2015.

Duan, J., Zeng, J., & Luo, B. (2014). Identification of Opinion Leaders Based on User Clustering and Sentiment Analysis. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01*

(pp. 377–383). IEEE Computer Society.

Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230.

Engel, J. F., Blackwell, R. D., & Kollat, D. J. (1978). Consumer Behavior, 3rd. *Holt, Rinehart andWinston Inc., New York*.

Everett, M. G., & Borgatti, S. P. (1999). The centrality of groups and classes. *The Journal of Mathematical Sociology*, *23*(3), 181–201.

Fan, W., & Gordon, M. D. (2014). The Power of Social Media Analytics. *Communications of the ACM*, *57*(6), 74–81. http://doi.org/10.1145/2602574

Feick, L. F., & Price, L. L. (1987). The market maven: A diffuser of marketplace information. *The Journal of Marketing*, 83–97.

Fletcher, K. (1988). An Investigation into the nature of problem recognition and deliberation in buyer behaviour. *European Journal of Marketing*, *22*(5), 58–66.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35–41.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239.

Geoffrion, A. M. (1976). The purpose of mathematical programming is insight, not numbers. *Interfaces*, *7*(1), 81–92.

Gerick, B. (2014). Oscars 2014: Ellen DeGeneres' all-star selfie sets Twitter record for most retweets. *New York Daily News*.

Goh, K.-Y., Heng, C.-S., & Lin, Z. (2013). Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content. *Information Systems Research*, *24*(1), 88–107.

Goodman, M. B., Booth, N., & Matic, J. A. (2011). Mapping and leveraging influencers in social media to shape corporate brand perceptions. *Corporate Communications: An International Journal*, *16*(3), 184–191.

Gray, P. H., Parise, S., & Iyer, B. (2011). Innovation Impacts of Using Social Bookmarking Systems. *MIS Quarterly*, *35*(3), 629–643.

Grönroos, C., & Voima, P. (2013). Critical service logic: making sense of value creation and co-creation. *Journal of the Academy of Marketing Science*, *41*(2), 133–150.

Grubmüller, V., Götsch, K., & Krieger, B. (2013). Social media analytics for future oriented policy making. *European Journal of Futures Research*, *1*(1), 20.

Grubmüller, V., Krieger, B., & Götsch, K. (2013). Social media analytics for government in the light of legal and ethical challenges. In *Proceedings of the 2013 International Conference for E-Democracy and Open Government 2013* (p. 185). Krems an der Donau, Austria.

Gualtieri, M., & Curran, R. (2014). The Forrester Wave: Big Data Streaming Analytics Platforms, Q3 2014 [Report]. Retrieved December 30, 2015, from Forrester Research: https://www.forrester.com/The+Forrester+Wave+Big+Data+Streaming+Analytics+Platforms+Q3+2014/fulltext/-/E-RES113442

Hajian, B., & White, T. (2011). Modelling influence in a social network: Metrics and evaluation.

In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (pp. 497–500). IEEE.

Han, J., Haihong, E., Le, G., & Du, J. (2011). Survey on NoSQL database. In *Pervasive computing and applications (ICPCA), 2011 6th international conference on* (pp. 363–366). IEEE.

Haveliwala, T. H. (2002). Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web* (pp. 517–526). ACM.

He, W., Wu, H., Yan, G., Akula, V., & Shen, J. (2015). A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, *52*(7), 801–812. http://doi.org/10.1016/j.im.2015.04.006

Heinrichs, J. H., & Lim, J. (2005). Model for organizational knowledge creation and strategic use of information. *Journal of the American Society for Information Science and Technology*, *56*(6), 620–629.

Helfat, C. E., Finkelstein, S., Mitchell, W., Peteraf, M., Singh, H., Teece, D., & Winter, S. G. (2009). *Dynamic capabilities: Understanding strategic change in organizations*. John Wiley & Sons.

Hennig-Thurau, T., Walsh, G., & Walsh, G. (2003). Electronic word-of-mouth: Motives for and consequences of reading customer articulations on the Internet. *International Journal of Electronic Commerce*, *8*(2), 51–74.

Hill, S., & Ready-Campbell, N. (2011). Expert stock picker: The wisdom of (experts in) crowds. *International Journal of Electronic Commerce*, *15*(3), 73–102.

Hoffman, D. D. L. D. L., & Fodor, M. (2010). Can You Measure the ROI of Your Social Media Marketing? *MIT Sloan Management Review*, *52*(1), 41–49. Retrieved from http://www.mitsmr-ezine.com/mitsmriphone11/fall2010/m2/MobileArticle.action?articleId=23732&mobileWeb=true&lm=1285614348000\nhttp://www.emarketingtravel.net/resources/can you mesur the ROI of your Social media marketing.pdf\nhttp://sloanreview.mit

Honey, C., & Herring, S. C. (2009). Beyond microblogging: Conversation and collaboration via Twitter. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on* (pp. 1–10). IEEE.

Hu, N., Sian, K. N., & Reddy, S. K. (2014). Ratings Lead You To The Product, Reviews Help You Clinch It? The Mediating Role of Online Review Sentiments on Product Sales. *Decision Support Systems*, *57*, 42–53.

Huber, G. P., & Daft, R. L. (1987). The information environments of organizations.

Hudli, S., Hudli, A., & Hudli, A. V. (2012). Identifying online opinion leaders using K-means clustering. In *Intelligent Systems Design and Applications (ISDA), 2012 12th International Conference on* (pp. 416–419). IEEE.

Huffaker, D. (2010). Dimensions of leadership and social influence in online communities. *Human Communication Research*, *36*(4), 593–617. http://doi.org/10.1111/j.1468-2958.2010.01390.x

Insight. (2016).

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR)*, *31*(3), 264–323.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (pp. 56–65). ACM.

Jiang, L., Ge, B., Xiao, W., & Gao, M. (2013). BBS opinion leader mining based on an improved PageRank algorithm using MapReduce. In *Chinese Automation Congress (CAC), 2013 IEEE Conference on* (pp. 392–396). IEEE.

Jing, L., & Lizhen, X. (2014). Identification of Microblog Opinion Leader Based on User Feature and Interaction Network. In *Web Information System and Application Conference (WISA)* (pp. 125–130). IEEE.

Kane, G. C., Alavi, M., Labianca, G. J., & Borgatti, S. (2012). What's different about social media networks? A framework and research agenda. *MIS Quarterly, Forthcoming*.

Kane, G. C., Alavi, M., Labianca, G. J., & Borgatti, S. (2014). What's different about social media networks? A framework and research agenda. *MIS Quarterly*, *38*(1), 274–304.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68.

Katz, E., & Lazarsfeld, P. F. (1955). *Personal Influence, The part played by people in the flow of mass communications*. Transaction Publishers.

Khan, F. H., Bashir, S., & Qamar, U. (2014). TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, *57*, 245–257.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, *54*(3), 241–251.

King, C. W., & Summers, J. O. (1970). Overlap of opinion leadership across consumer product categories. *Journal of Marketing Research*, 43–50.

Klein, G., Pliske, R., Crandall, B., & Woods, D. D. (2005). Problem detection. *Cognition, Technology & Work*, *7*(1), 14–28.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, *46*(5), 604–632.

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. S. (1999). The web as a graph: measurements, models, and methods. In *Computing and combinatorics* (pp. 1–17). Springer.

Kumar, S., Morstatter, F., & Liu, H. (2014). Storing Twitter Data. In *Twitter Data Analytics* (pp. 23–33). Springer.

Kurniawati, K., Shanks, G., & Bekmamedova, N. (2013). The Business Impact Of Social Media Analytics. In *Proceedings of the European Conference on Information Systems* (p. 48). Utrecht, The Netherlands.

Langner, S., Hennigs, N., & Wiedmann, K.-P. (2013). Social persuasion: targeting social identities through social influencers. *Journal of Consumer Marketing*, *30*(1), 31–49.

Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1944). *The People's Choice: How the Voter Makes*

Up His Mind in a Presidential Campaign. Duell, Sloan and Pearce.

Leavitt, N. (2010). Will NoSQL databases live up to their promise? *Computer*, *43*(2), 12–14.

Li, B., Wong, K., Zhou, L., Wei, Z., & Xu, J. (2013). Pests Hidden in Your Fans: An Effective Approach for Opinion Leader Discovery. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data* (pp. 227–237). Springer.

Li, F., & Du, T. C. (2011). Who is talking? An ontology-based opinion leader identification framework for word-of-mouth marketing in online social blogs. *Decision Support Systems*, *51*(1), 190–197.

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, *69*, 14–23.

Libai, B., Bolton, R., Bügel, M. S., De Ruyter, K., Götz, O., Risselada, H., & Stephen, A. T. (2010). Customer-to-customer interactions: broadening the scope of word of mouth research. *Journal of Service Research*, *13*(3), 267–282.

Lin, C.-L., & Kao, H.-Y. (2010). Blog popularity mining using social interconnection analysis. *IEEE Internet Computing*, (4), 41–49.

Liu, H., Yu, X., & Lu, J. (2013). Identifying TOP-N opinion leaders on local social network. In *Smart and Sustainable City 2013 (ICSSC 2013), IET International Conference on* (pp. 325–328). IET.

Ma, N., & Liu, Y. (2014). SuperedgeRank algorithm and its application in identifying opinion leader of online public opinion supernetwork. *Expert Systems with Applications*, *41*(4), 1357–1368.

Mangold, W. G., & Faulds, D. J. (2009). Social Media: The new hybrid element of the promotion mix. *Business Horizons*, *52*(4), 357–365. http://doi.org/10.1016/j.bushor.2009.03.002

Marett, K., & Joshi, K. D. (2009). The decision to share information and rumors: Examining the role of motivation in an online discussion forum. *Communications of the Association for Information Systems*, *24*(1), 4.

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia* (pp. 31–40). ACM.

Mayeh, M., Scheepers, R., & Valos, M. (2012). Understanding the role of social media monitoring in generating external intelligence. In *Proceedings of the 2012 23rd Australasian Conference on Information Systems* (pp. 1–10). Geelong, Australia.

Melville, P., Sindhwani, V., & Lawrence, R. (2009). Social media analytics: Channeling the power of the blogosphere for marketing insight. In *Proceedings of the 2009 1st Workshop on Information in Networks (WIN 2009)*. Manhattan, NY, USA.

Minocha, S., & Roberts, D. (2008). Social, usability, and pedagogical factors influencing students' learning experiences with wikis and blogs. *Pragmatics & Cognition*, *16*(2), 272–306.

Mosley, R. C. (2012). Social Media Analytics: Data Mining Applied to Insurance Twitter Posts. In *Proceedings of the 2012 Casualty Actuarial Society E-Forum* (Vol. 2, pp. 1–36). Arlington, Virginia, USA.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful review? A study of customer reviews on Amazon. com. *MIS Quarterly*, *34*(1), 185–200.

Munzel, A., & H. Kunz, W. (2014). Creators, multipliers, and lurkers: who contributes and who benefits at online review sites. *Journal of Service Management*, *25*(1), 49–74.

Nagurney, A., & Dong, J. (2002). *Supernetworks: decision-making for the information age*. Elgar, Edward Publishing, Incorporated.

Nagurney, A., & Wakolbinger, T. (2005). Supernetworks: An introduction to the concept and its applications with a specific focus on knowledge supernetworks. *International Journal of Knowledge, Culture and Change Management*, *4*, 1523–1530.

Ngai, E. W. T., Tao, S. S. C., & Moon, K. K. L. (2015). Social media research: Theories, constructs, and conceptual frameworks. *International Journal of Information Management*, *35*(1), 33–44. http://doi.org/10.1016/j.ijinfomgt.2014.09.004

Nutt, P. C. (2007). Intelligence gathering for decision making. *Omega*, *35*(5), 604–622.

O'connor, G. C., & Rice, M. P. (2001). Opportunity recognition and breakthrough innovation in large established firms. *California Management Review*, *43*(2), 95–116.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, *11*, 169–198.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: bringing order to the Web.

Raymond, E. (1999). The cathedral and the bazaar. *Knowledge, Technology & Policy*, *12*(3), 23–49.

Regan, K. (2015). 10 Amazing Social Media Growth Stats From 2015. Retrieved from http://www.socialmediatoday.com/social-networks/kadie-regan/2015-08-10/10-amazing-social-media-growth-stats-2015

Ribarsky, W., Xiaoyu Wang, D., & Dou, W. (2013). Social Media Analytics for Competitive Advantage. *Computers & Graphics*, *38*, 328–331.

Ring, P. S., & Rands, G. P. (1989). Sensemaking, understanding, and committing: Emergent interpersonal transaction processes in the evolution of 3M's microgravity research program. *Research on the Management of Innovation: The Minnesota Studies*, 337–366.

Rodgers, K., & Scobie, W. (2015). Sealfies, seals and celebs: expressions of Inuit resilience in the Twitter era. *Interface*, *7*(1), 70–97.

Rogers, E. M., & Cartano, D. G. (1962). Methods of measuring opinion leadership. *Public Opinion Quarterly*, *26*(3), 435.

Russom, P. (2011). Big data analytics. *TDWI Best Practices Report, Fourth Quarter*, 1–35.

SAP. (2014). Continuous Intelligence with Event Stream Processing [White Paper]. Retrieved December 31, 2015, from SAP Community Network:: http://www.sdn.sap.com/irj/scn/go/portal/prtroot/docs/library/uuid/e047a3db-45f1-3010-dea3-e5875b2a663b?QuickLink=index&overridelayout=true&59575491391812

Savolainen, R. (1993). The sense-making theory: Reviewing the interests of a user-centered approach to information seeking and use. *Information Processing & Management*, *29*(1), 13–28.

Schaust, S., Walther, M., & Kaisser, M. (2013). Avalanche: Prepare, Manage, and Understand Crisis Situations Using Social Media Analytics. In *Proceedings of the 2013 10th*

International ISCRAM Conference (pp. 852–857).

Schenk, M., & Rössler, P. (1997). The rediscovery of opinion leaders. An application of the personality strength scale. *Communications*, *22*(1), 5–30.

Sheffi, Y. (1985). Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods. Prentice-Hall, Englewood Cliffs, NJ.

Simon, H. A. (1960). The new science of management decision.

Simon, H. A. (1996). *The sciences of the artificial*. MIT press.

Sinha, V., Subramanian, K. S., Bhattacharya, S., & Chaudhary, K. (2012). The contemporary framework on social media analytics as an emerging tool for behavior informatics, HR analytics and business process. *Journal of Contemporary Management Issues*, *17*(2), 65–84.

Song, K., Wang, D., Feng, S., & Yu, G. (2011). Detecting opinion leader dynamically in chinese news comments. In *Web-Age Information Management* (pp. 197–209). Springer.

Stapleton, J. J. (2003). *Executive's guide to knowledge management: the last competitive advantage*. John Wiley & Sons.

Starbuck, W. H., & Milliken, F. J. (1988). Executives' perceptual filters: What they notice and how they make sense.

Steiger, D. M. (1998). Enhancing user understanding in a decision support system: a theoretical basis and framework. *Journal of Management Information Systems*, *15*(2), 199–220.

Sterne, J. (2010). *Social media metrics: How to measure and optimize your marketing investment*. Wiley.

Stieglitz, S., & Dang-Xuan, L. (2012). Social media and political communication: a social media analytics framework. *Social Network Analysis and Mining*, 1–15. http://doi.org/10.1007/s13278-012-0079-3

Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social Media Analytics: An Interdisciplinary Approach and Its Implications for Information Systems. *Business & Information Systems Engineering*, *6*(2), 89–96.

Subramani, M. R., & Rajagopalan, B. (2003). Knowledge-sharing and influence in online social networks via viral marketing. *Communications of the ACM*, *46*(12), 300–307.

Summers, J. O. (1970). The identity of women's clothing fashion opinion leaders. *Journal of Marketing Research*, 178–185.

Susarla, A., Oh, J.-H., & Tan, Y. (2012). Social networks and the diffusion of user-generated content: Evidence from YouTube. *Information Systems Research*, *23*(1), 23–41.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307.

Teece, D. J. (2007). Explicating dynamic capabilities: the nature and microfoundations of (sustainable) enterprise performance. *Strategic Management Journal*, *28*(13), 1319–1350.

Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. *Strategic Management Journal*, *18*(7), 509–533.

Teo, T. S. H., & Choo, W. Y. (2001). Assessing the impact of using the Internet for competitive intelligence. *Information & Management*, *39*(1), 67–83.

Torres, J., Baquerizo, G., Vaca, C., & Pel, E. (2016). Characterizing influential leaders of Ecuador on Twitter using computational intelligence. In *2016 Third International Conference on eDemocracy & eGovernment (ICEDEG)* (pp. 159–163). IEEE.

Vorvoreanu, M., Boisvenue, G. A., Wojtalewicz, C. J., & Dietz, E. J. (2013). Social media marketing analytics: A case study of the public's perception of Indianapolis as Super Bowl XLVI host city. *Journal of Direct, Data and Digital Marketing Practice*, *14*(4), 321–328.

Wally, S., & Baum, J. R. (1994). Personal and structural determinants of the pace of strategic decision making. *Academy of Management Journal*, *37*(4), 932–956.

Wang, D., Irani, D., & Pu, C. (2013). A study on evolution of email spam over fifteen years. In *Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom), 2013 9th International Conference Conference on* (pp. 1–10). IEEE.

Wang, F. Y., Carley, K. M., Zeng, D., & Mao, W. (2007). Social Computing : From Social Informatics to Social Intelligence. *IEEE Computer Society*, *22*(2), 79–83.

Wang, G. A., Jiao, J., Abrahams, A. S., Fan, W., & Zhang, Z. (2013). ExpertRank: A topic-aware expert finding algorithm for online knowledge communities. *Decision Support Systems*, *54*(3), 1442–1541.

Wang, J.-W., Rong, L.-L., Deng, Q.-H., & Zhang, J.-Y. (2010). Evolving hypernetwork model. *The European Physical Journal B*, *77*(4), 493–498.

Wang, Q., Xu, J., Li, H., & Craswell, N. (2013). Regularized latent semantic indexing: A new approach to large-scale topic modeling. *ACM Transactions on Information Systems (TOIS)*, *31*(1), 5.

Wei, X., & Hong, H. J. (2013). Micro Blogging Opinion Leaders of Quality Circle Based on Social Network Analysis and Mining. *Journal of Applied Sciences*, *13*(18), 3698.

Weick, K. E. (1993). The collapse of sensemaking in organizations: The Mann Gulch disaster. *Administrative Science Quarterly*, 628–652.

Weick, K. E. (1995). *Sensemaking in organizations* (Vol. 3). Sage.

Weinberg, B. D., & Pehlivan, E. (2011). Social spending: Managing the social media mix. *Business Horizons*, *54*(3), 275–282. http://doi.org/10.1016/j.bushor.2011.01.008

White, D. R., & Borgatti, S. P. (1994). Betweenness centrality measures for directed graphs. *Social Networks*, *16*(4), 335–346.

White, J. S., Matthews, J. N., & Stacy, J. L. (2012). Coalmine: an experience in building a system for social media analytics. In *Proceedings of the 2012 SPIE 8408, Cyber Sensing*. International Society for Optics and Photonics.

Wootton, J. (2014). SAP HANA Smart Data Streaming: Technical Overview [Whtie Paper]. Retrieved December 30, 2015, from SAP Community Network: http://www.sdn.sap.com/irj/scn/go/portal/prtroot/docs/library/uuid/2093fccb-ee5d-3210-c3b0-e17cc7d684f2?QuickLink=index&overridelayout=true&59661390722454

Wu, I., & Lin, Y.-S. (2012). WNavi s: Navigating Wikipedia semantically with an SNA-based summarization technique. *Decision Support Systems*, *54*(1), 46–62.

Xiao, Y., & Xia, L. (2010). Understanding opinion leaders in bulletin board systems: Structures and algorithms. In *Local Computer Networks (LCN), 2010 IEEE 35th Conference on* (pp. 1062–1067). IEEE.

Yang, M., Kiang, M., Ku, Y., Chiu, C., & Li, Y. (2011). Social media analytics for radical opinion mining in hate group web forums. *Journal of Homeland Security and Emergency Management*, *8*(1).

Yi-si, C., & Guo-xin, L. (2012). Leading users and opinion leaders in social networks of university students. In *Management Science and Engineering (ICMSE), 2012 International Conference on* (pp. 89–94). IEEE.

Zeng, D., Chen, H., Lusch, R., & Li, S.-H. (2010). Social media analytics and intelligence. *Intelligent Systems, IEEE*, *25*(6), 13–16.

Zeng, X., & Wei, L. (2013). Social ties and user content generation: Evidence from Flickr. *Information Systems Research*, *24*(1), 71–87.

Zhai, Z., Xu, H., & Jia, P. (2008). Identifying opinion leaders in BBS. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on* (Vol. 3, pp. 398–401). IEEE.

Zhou, H., Zeng, D., & Zhang, C. (2009). Finding leaders from opinion networks. In *Intelligence and Security Informatics, 2009. ISI'09. IEEE International Conference on* (pp. 266–268). IEEE.

Zhou, X., Yang, J., Zhang, J., & Lin, Z. (2014). A BBS opinion leader mining algorithm based on topic model. *Journal of Computational Information Systems*, *10*(6), 2571–2578.

Zhu, Y.-Q., & Chen, H.-G. (2015). Social media and human need satisfaction: Implications for social media marketing. *Business Horizons*, *58*(3), 335–345.

Ziyi, L., Jing, C. F. S., Donghong, S., & Yongfeng, H. (2013). Research on methods to identify the opinion leaders in Internet community. In *Software Engineering and Service Science (ICSESS), 2013 4th IEEE International Conference on* (pp. 934–937). IEEE.

Zwass, V. (2010). Co-creation: Toward a taxonomy and an integrated research perspective. *International Journal of Electronic Commerce*, *15*(1), 11–48.

Vita

1. Place of Birth: Taipei, Taiwan

2. Education

Chung Yuan Christian University	Zhongli, Taiwan
M.B.A., Information Management	May 2007
Chung Yuan Christian University	Zhongli, Taiwan
B.A., Information Management	May 2005

3. Publications
   - Y. Wang, S. H. Hsiao, Z. Yang, & N. Hajli, N. (2015). The impact of sellers' social influence on the co-creation of innovation with customers and brand awareness in online communities. *Industrial Marketing Management*
   - Liang Chen, Clyde W. Holsapple, S. H. Hsiao, Zhihong Ke, J. Y. Oh, & Zhiguo Yang, forthcoming, *Journal of the American Society for Information Science and Technology*
   - Clyde W. Holsapple, S. H. Hsiao, & Y. Y. Oh, Parameters of Knowledge Management Success, with Clyde W. Holsapple & J. Y. Oh, forthcoming, *Successes and Failures of Knowledge Management*, Jay Liebowitz (Ed.), Morgan Kaufmann/Elsevier.

4. Conference Presentations
   - S. H. Hsiao & Ram Pakath. (2015). Who Are the Opinion Leaders? A Relative Assessment of Opinion Leader Mining Algorithms. 2015 *INFORMS Annual Meeting*. Philadelphia, PA.
   - S. H. Hsiao, Yichuan Wang, Zhiguo Yang, & Nick Hajli. (2015). Leveraging Co-innovation Practices on Business-to-Business Virtual Communities. *21th Annual Americas Conference on Information Systems (AMCIS).* Fajardo, Puerto Rico.
   - S. H. Hsiao & Yichuan Wang. (2015). The Effect of Social Factors on User-Generated Content Productivity: Evidence from Flickr.com. *21th Annual Americas Conference on Information Systems (AMCIS).* Fajardo, Puerto Rico.
   - Y. Y. Wang & S. H. Hsiao. (2014). IT-enabled Intangibles and IT Capabilities: A Study from the Resource-based view and IS Strategy Perspective. *20th Annual Americas Conference on Information Systems (AMCIS)*. Savannah, GA.
   - C. W. Holsapple, S. H. Hsiao, & Ram Pakath. (2014) Business Social Media Analytics: Definition, Benefits, and Challenges. *20th Annual Americas Conference on Information Systems (AMCIS)*. Savannah, GA.
   - C. W. Holsapple, S. H. Hsiao, & Ram Pakath. (2014). Business Social Media Analytics: Definition, Benefits, Challenges, and a Conceptual Model. *45th Decision Sciences Institute Annual Meeting.* Tampa, FL.
   - S. H. Hsiao & Anita Lee-Post. (2013). Co-Creation and Competitiveness: a PAIR Perspective. *44th Decision Sciences Institute Annual Meeting.* Baltimore, MD.

5. Teaching Experience

Main Instructor

Undergraduate Level *Spring 2013*

Information Systems in the Modern Enterprise (analytics major required)

AN325, University of Kentucky (Rating: 3.6/ 4.0)

Teaching Assistant

Graduate Level

Quantitative Analysis in Business Decision Making *Spring 2015*

DIS651, Evening MBA course, University of Kentucky

Quantitative Analysis in Business Decision Making *Summer 2(*

DIS651, MBA course, University of Kentucky

Undergraduate Level

Analyzing Business Operation *Fall 2011– Fall 2015*

AN300, University of Kentucky

Data Mining *Fall 2014*

AN420G, University of Kentucky

6. Work Experience

University of Kentucky *Lexington, Kentucky*

Research Assistant *2013- 2015*

University of Kentucky *Lexington, Kentucky*

Teaching Assistant *2011- 2015*

Office of International Affairs, National Yang-Ming University *Taipei, Taiwan*

IT Support Specialist *2010- 2011*

Institute for Information Industry *Taipei, Taiwan*

Course Assistant *2009- 2010*

The Republic of China Army *Taipei, Taiwan*

Corporal *2008- 2009*

Campus Youth E-service Volunteer Center, Ministry of Education *Taipei, Taiwan*

System Engineer *2006-2007*

7. Awards and Honors

Max Steckler Fellowship *2015*

Gatton College Doctoral Fellowship *2011- 2014*

ING Antai National MBA Thesis Award, Taiwan *2007*

National Youth Public Participation Award *2007*