

AN EXPLORATORY ANALYSIS OF TWITTER  
KEYWORD-HASHTAG NETWORKS  
AND KNOWLEDGE DISCOVERY APPLICATIONS

A Dissertation Presented

by

Ahmed Abdeen Hamed

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements  
for the Degree of Doctor of Philosophy  
Specializing in Computer Science

May, 2014

Accepted by the Faculty of the Graduate College, The University of Vermont, in partial fulfillment of the requirements for the degree of Doctor of Philosophy, specializing in Computer Science.

Dissertation Examination Committee:

Advisor

---

Xindong Wu, Ph.D.

---

Josh Bongard, Ph.D.

---

Alan Rubin, Ph.D.

Chairperson

---

Stephen Higgins, Ph.D.

Dean, Graduate College

---

Cynthia J. Forehand, Ph.D.

Date: March 27, 2014

## Abstract

The emergence of social media has impacted the way people think, communicate, behave, learn, and conduct research. In recent years, a large number of studies have analyzed and modeled this social phenomena. Driven by commercial and social interests, social media has become an attractive subject for researchers. Accordingly, new models, algorithms, and applications to address specific domains and solve distinct problems have erupted. In this thesis, we propose a novel network model and a path mining algorithm called *HashnetMiner* to discover implicit knowledge that is not easily exposed using other network models. Our experiments using *HashnetMiner* have demonstrated anecdotal evidence of drug-drug interactions when applied to a drug reaction context.

The proposed research comprises three parts built upon the common theme of utilizing hashtags in tweets.

1. Digital Recruitment on Twitter. We build an expert system shell for two different studies: (1) a nicotine patch study where the system reads streams of tweets in real time and decides whether to recruit the senders to participate in the study, and (2) an environmental health study where the system identifies individuals who can participate in a survey using Twitter.
2. Does Social Media Big Data Make the World Smaller? This work provides an exploratory analysis of large-scale keyword-hashtag networks (K-H) generated from Twitter. We use two different measures, (1) the number of vertices that connect any two keywords, and (2) the eccentricity of keyword vertices, a well-known centrality and shortest path measure. Our analysis shows that K-H networks conform to the phenomenon of the shrinking world and expose hidden paths among concepts.
3. We pose the following biomedical web science question: Can patterns identified in Twitter hashtags provide clinicians with a powerful tool to extrapolate a new medical therapies and/or drugs? We present a systematic network mining method *HashnetMiner*, that operates on networks of medical concepts and hashtags. To the best of our knowledge, this is the first effort to present Biomedical Web Science models and algorithms that address such a question by means of data mining and knowledge discovery using hashtag-based networks.

## Citations

Material from this dissertation has been published in the following form:

Hamed, Ahmed Abdeen and Wu, Xindong and Fingar, James.. (2013). A Twitter-based Smoking Cessation Recruitment System. *ASONAM '13 Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Pages 854-861.

Hamed, Ahmed Abdeen and Wu, Xindong and Rubin, Alan.. (2014). A Twitter Recruitment Intelligent System: Association Rule Mining for Smoking Cessation. *Journal of Social Network Analysis and Mining*, invited submission.

AND

Hamed, Ahmed Abdeen and Wu, Xindong.. (2014). Does Social Media Big Data Make the World Smaller? An Exploratory Analysis of Keyword-Hashtag Networks. *IEEE International Congress on Big Data – BigData'14*, accepted, to appear.

AND

Hamed, Ahmed Abdeen and Wu, Xindong and Fandy, Tamer and Lee, Byung Suk.. (2014). Biomedical Web Science: Mining Patterns in Concept-Hashtag Networks. *ACM Web Science 2014 Conference*, submitted.

## **Dedication**

*To my parents: Aisha Ouda and Abdel Raouf Abdeen Hamed  
To family members in Egypt  
To my daughter: Laila M. Hamed*

## Acknowledgements

I would like thank my primary Advisor, Dr. Xindong Wu, the one who introduced me to Artificial Intelligence and Data Mining. He believed in me and my own ideas and give wings to them. It is a true delight to receive my academic training by such a wonderful scientist and scholar. I would like also to thank Dr. Josh Bongard for his role on the defense committee I would like to thank Dr. Alan Rubin, the one whom without whose help, I would have never been able to find my way back into grad-school. His support and guidance can never be forgotten. I would like to thank Dr. Stephen Higgins for playing the role of the chair and being a big fan of my work and research.

I would like to acknowledge both Dr. Benjamin Littenberg and Dr. John Hughes for offering scholarships and supporting my academic career.

I would like to thank both Dr. Tamer Fandy and Dr. Glen Myer for their feedback on the pharmacology knowledge discovery component of my research.

I would like to thank all my UVM CS faculty and staff whom I learned so much from: Dr. Donna Rizzo, Dr. Peter Dodds, Dr. Bill Jefferys, Alison Pechenick, Robert Erickson, and Penny French.

I would like to acknowledge the technical architects of the VACC and CEMS IT. Particularly Jim Lawson and Victor Rossi for accommodating the various needs for my development and experimental environment.

I would like to thank my fellow CS graduate students for their ultimate support through out this journey: Dr. Haiguang Li, Mark Wagy, Sepehr Mohammadian, and Afsoon Yz.

I would like to thank Dr. Ramiro Barrantes-Reynolds, and Dr. Carolyn Bonifield for their continuous support and the valuable help and discussions

I would like thank my former advisors of Indiana University, Dr. Amr Sabry, Dr. Fil Menczer, Dr. Luis Rocha, Dr. Haixu Tang, Dr. Dirk Van Gucht, Dr. Steven Johnson, Dr. Daniel Leivant, Dr. Ahmed YoussefAgha, Adrian German, and Dr. David Lohrmann.

I would like to thank the open-source community: Twitter4J, Weka, iGraph, Jung, Gephi, GIS Tools, and Google Analytics.

I would like to thank Dr. Dr. Tamás Nepusz of the Python IGraph community for the valuable discussions. I would like to acknowledge her role for answering questions related to network visualization and the insights she provided to present all the networks visualizations of this thesis. I would also like to acknowledge Dr. Lada Adamic for the Social Network Analysis course she offered on Coursera.

Finally, I would like acknowledge the support received from Lindsay Vannarsdall and for helping with the overall proof-reading this thesis.

# Table of Contents

<b>Citations</b> . . . . .	<b>ii</b>
<b>Dedication</b> . . . . .	<b>iii</b>
<b>Acknowledgements</b> . . . . .	<b>iv</b>
<b>List of Figures</b> . . . . .	<b>ix</b>
<b>List of Tables</b> . . . . .	<b>x</b>
<b>1 Background and Significance</b>	<b>2</b>
<b>1.1 Motivation</b> . . . . .	<b>2</b>
<b>1.2 Background</b> . . . . .	<b>4</b>
1.2.1 Data Mining and Knowledge Discovery . . . . .	5
1.2.2 Mining Tweets . . . . .	5
1.2.3 Network Pattern Mining . . . . .	6
<b>1.3 Contributions</b> . . . . .	<b>6</b>
<b>1.4 Thesis Outline</b> . . . . .	<b>7</b>
<b>2 A Rule-based Twitter Recruitment Intelligent System</b>	<b>9</b>
<b>2.1 Introduction</b> . . . . .	<b>10</b>
2.1.1 Recruitment Strategy and Specification . . . . .	11
2.1.2 Contributions . . . . .	12
<b>2.2 Recruiting Twitter Users</b> . . . . .	<b>13</b>
2.2.1 Problem Formulation . . . . .	13
2.2.2 Proposed Approach . . . . .	13
Twitter Logistics . . . . .	14
2.2.3 System Architecture . . . . .	16
Twitter Monitors . . . . .	16
Tweet Analyzer . . . . .	16
Database of Tweets, Hashtags, Web Links . . . . .	17
Event Processor . . . . .	18
<b>2.3 System Performance And Evaluation</b> . . . . .	<b>19</b>



<b>2.4 Twitter Recruitment Algorithm</b>	<b>22</b>
2.4.1 Non-Textual Feature Scoring Algorithm	22
PageRank Theoretical Background	24
2.4.2 Rule-based System	25
Tweet Rule	26
Reply Rule	26
Retweet Rule	27
<b>2.5 Association Rule Mining Knowledge Acquisition Using Apriori Algorithm</b>	<b>28</b>
2.5.1 Data Description and Gathering	28
2.5.2 Setting Up Minimum Support	29
2.5.3 Experimental Results	30
<b>2.6 Discussions</b>	<b>31</b>
<b>2.7 Acknowledgments</b>	<b>34</b>
<b>3 An Exploratory Analysis of K-H Networks</b>	<b>35</b>
<b>3.1 Motivation</b>	<b>36</b>
<b>3.2 Related Work</b>	<b>38</b>
<b>3.3 Definitions</b>	<b>40</b>
<b>3.4 Data Collection, Modeling and Network Construction</b>	<b>41</b>
<b>3.5 Experiments</b>	<b>44</b>
3.5.1 Small dataset experiments	44
3.5.2 Large dataset experiments	48
<b>3.6 Discussion</b>	<b>51</b>
<b>3.7 Conclusions</b>	<b>53</b>
<b>3.8 Acknowledgments</b>	<b>54</b>
<b>4 Biomedical Web Science: Mining Patterns in Concept-Hashtag Networks</b>	<b>55</b>
<b>4.1 Introduction</b>	<b>56</b>
<b>4.2 Data</b>	<b>59</b>

<b>4.3 Preliminary</b> . . . . .	<b>60</b>
4.3.1 Proof of Concept . . . . .	62
4.3.2 Emergent Hashtags Rule Discovery . . . . .	63
4.3.3 Types of Hashtags Found . . . . .	64
<b>4.4 Method</b> . . . . .	<b>67</b>
4.4.1 Association Network Construction . . . . .	67
4.4.2 Network Mining Heuristics . . . . .	68
4.4.3 Mapping to MeSH and Linking to PubMed . . . . .	71
<b>4.5 Analysis</b> . . . . .	<b>72</b>
4.5.1 Browsing Findings using a WebClient . . . . .	75
<b>4.6 Conclusions</b> . . . . .	<b>76</b>
<b>4.7 Acknowledgments</b> . . . . .	<b>80</b>
<b>5 Conclusions and Future Work</b>	<b>81</b>
<b>5.1 Conclusions</b> . . . . .	<b>81</b>
<b>5.2 Future Work</b> . . . . .	<b>83</b>
<b>5.3 Acknowledgments</b> . . . . .	<b>85</b>
<b>References</b>	<b>86</b>

## List of Figures

2.1	Twitter Recruitment System Architecture . . . . .	19
2.2	TobaccoQuit Wordcloud Visualization . . . . .	20
2.3	TobaccoQuit Traffic using BitLy . . . . .	21
2.4	Association Rule Mining using Apriori in Weka . . . . .	29
2.5	KK vs KH dataset 1 experiments with three minsup levels showing the gain ratio favoring KH . . . . .	31
2.6	KK vs KH dataset 2 experiments with three minsup levels showing the gain ratio favoring KH . . . . .	32
3.1	Four sets of keywords used for small datasets. . . . .	43
3.2	K-K networks of the four domains. . . . .	45
3.3	K-H networks of the four domains. . . . .	46
3.4	Some of the keywords connected in the K-H networks while not in the K-K networks. . . . .	47
3.5	The four-domain K-H network. . . . .	48
3.6	The four-domain K-K network. . . . .	49
3.7	Eccentricity Percentiles . . . . .	50
3.8	Vertex eccentricity measured from using Erdős and Rényi’s random network model. . . . .	51
4.1	A WordCloud list of drug brand names used to search Twitter Streaming APIs . . . . .	60
4.2	Impact of Hashtags When Introduced to Disconnected Networks . . . . .	64
4.3	Association Rules mined by running Apriori against a small dataset of 25 tweets . . . . .	65
4.4	A, B, C Nodes form an Open Triad, and Possible Closures Using Hashtags or MeSH Terms Combinations . . . . .	69
4.5	Impact of Hashtags When Introduced to Disconnected Networks. Top figure shows the drug associations when no hashtags are absent. Bottom figure shows closed patterns when hashtags are incorporated . . . . .	71
4.6	PubMed Mesh-Hashtag Linking Web Application . . . . .	77
4.7	Marijuana, Ibuprofen and Alzheimer on Twitter . . . . .	80

## List of Tables

2.1	Pre-prepared Recruitment Sample Tweets . . . . .	18
2.2	TobaccoQuit Account Statistics . . . . .	20
2.3	Experiments settings . . . . .	30
3.1	Various Keyword-Hashtag Patterns and Their Corresponding Frequencies . . . . .	50
4.1	Dataset2 – Real-World tweets . . . . .	61
4.2	Transaction Types and Samples. Records with “?” Indicate Missing Values . . . . .	63
4.3	Mapping Drug to Generics and Linking to MH and EC/RN . . . . .	72
4.4	Rules found when searching for disconnected vertices pattern and open triad patterns	74



# Chapter 1

## Background and Significance

### 1.1 Motivation

Several hundred years ago, a great English mathematician named Isaac Newton was sitting under a tree. He saw an apple falling from the tree and he realized that some force must act on the apple to cause it to fall (Conner, 2008). Newton called this new force “gravity” and reasoned that gravitational forces act upon all objects on the earth (D’Amico, 2009). The fact that apples fall from trees, that objects fall toward the earth rather than floating up toward the sky, was an obvious, unexamined phenomenon. However, Newton did examine this phenomenon and observed it closely. His observations led to the development of the gravity theory (Clegg, 2012). This discovery is considered an essential milestone in the thinking and research that eventually led to space travel in the 20th century. (Linehan, 2011; Ferreira & Starkmann, 2009).

A reader of this dissertation may think: Twitter and Facebook enabled hashtags; therefore, hashtags must be useful. A question a reader may ask is: do we really need an entire dissertation to state the obvious? Just as Newton was curious to prove that gravity exists and founded the basis of gravity theory, we also believe that we need to examine hashtags to determine whether they are useful in some way. How are they useful to human knowledge? To what extent? Newton’s unconventional questioning of a simple event in the world around him still inspires scientists today to examine the

“obvious” world and to tease out how their observations might apply broadly. This dissertation was inspired and sustained by that spirit, though not purporting to make the kind of impact Newton made. (One can hope, though.)

Following the invention of the Internet, the world of information has expanded. People are able to connect from many parts of the world to share information via electronic mail (email) and blogging sites (Castells & Cardoso, 2006). Social media, an application built on the Internet, has created an environment for individuals to interact and cooperate across vast distances (Hanaki et al., 2007). Specifically, social media has enabled people to connect to others through sharing thoughts, images, articles, and websites (Abel et al., 2009), (Kaplan & Haenlein, 2010), and (Mazman & Usluel, 2010).

Social media sites generate seemingly unlimited types of knowledge and open a new research domain for social scientists, physicists, computer scientists, mathematicians, biologists, and medical scientists (McCarthy et al., 2013). Twitter, one of the best known social media outlets, has attracted the attention of scientists more than any other social media. The various characteristics of Twitter feeds, also known as tweets, make Twitter an appealing vehicle for conducting research. A tweet may contain words, links to external websites, an image, and/or a special device called a “hashtag”. Additionally, a tweet cannot contain more than 140 characters. Due to this limitation, people have found interesting ways to use hashtags to overcome the length limit. Among the various usages, a hashtag can be used to express an idea (e.g., #SmallBusiness), to invite people to an event (e.g., #networking), or to recommend a product (e.g., #eCig). Moreover, people have personalized hashtags to express how they feel (e.g., #excited), what they are doing (e.g., #coffee, #cig), and where they travel (e.g., #Paris, #NiagaraFalls).

The emergence of such hashtags on Twitter marks the beginning of an era full of unexpected discoveries. Hashtags are no longer just a simple sequence of characters to label a tweet with a particular topic. The past few years have seen an ever-evolving usage of hashtags that encompasses a wide range of knowledge. Each individual hashtag functions similarly to a neuron in the human

brain, sending a specific signal and performing a distinct task. Among a large spectrum of signals, one may point to a location (#US), an event (#TobaccoControl, #HealthcareReform, #PotLegalization), outbreaks (#HIN1, #Norovirus), day of the week (#TGIF), or the name of a significant person (#MichelleObama, #Oprah). Hashtags play a crucial role for disseminating and connecting nuggets of information. Gaining a good understanding of how hashtags function is bound to reveal a wealth of unexpected knowledge. However, studying hashtag semantics is a very difficult task. It takes human intelligence to assess whether a single hashtag is significant, its interpretations and meanings. The sheer volume of hashtags makes it impossible to make these evaluations without automating the task using computational methods and algorithms.

Developing a computational method to identify significant hashtags requires a great deal of creativity. Such a method must be comparable to establishing sound and measurable theories. Consequently, the study of hashtags has proven to be appealing to scientists in recent years. Lehmann et al. (Lehmann et al., 2012) focused their study on hashtag peaks, which lead them to discover spikes of collective attention on Twitter. When popular hashtags are identified, they are linked back to their original tweets for further analysis. Romero et al. studied the hashtag spread on a network defined by the interactions among Twitter users. The study found significant variation of how widely-used hashtags spread on Twitter (Romero et al., 2011). Conover et al. (Conover et al., 2011) studied hashtags as an important feature for clustering political polarization on Twitter. Wang et al. (Wang et al., 2013) proposed a model that analyzes the traffic patterns of the hashtags gathered from streaming tweets to generate adaptive subsequent queries.

## **1.2 Background**

This dissertation is based upon three cornerstone areas: (1) data mining and knowledge discovery, (2) mining data stream in general and Tweets in particular, and (3) network pattern mining algorithms. The following is some background on the most recent work in each area.



### **1.2.1 Data Mining and Knowledge Discovery**

Data Mining is one of the fundamentals of this dissertation. Due to the massive number of references in this area, we present a survey paper which discusses the top-10 data mining algorithms (Wu et al., 2007). This reference is one of the most comprehensive reference documents to date, documenting significant algorithms. It covers all data mining branches which involve around cluster learning and analysis (e.g., K-Means (MacQueen, 1967), EM (McLachlan & Peel, 2000)), classification (e.g., C.45 (Quinlan, 1993), Naive Bayes (Hand & Yu, 2001), CART (Breiman et al., 1984), SVM (Vapnik, 1995), kNN (Hastie & Tibshirani, 1996)) and association analysis (Apriori (Agrawal & Srikant, 1994)), link analysis (PageRank (Brin & Page, 1998)) and bagging and boosting (AdaBoost (Freund & Schapire, 1995)).

### **1.2.2 Mining Tweets**

The literature on tweet mining and analysis is very rich. Here we list some of the most related and recent work in this area: Yang et al. (X. Yang et al., 2012) presented a framework for analyzing and summarizing Twitter feeds. Additionally, Meng et al. (Meng et al., 2012), presented an entity-centric, topic-oriented opinion summarization in Twitter to solve the same summarization problem. Yang et al. (M. Yang et al., 2012) also analyzed tweet posts in order to make a decision about whether to retweet a post based on its interestingness. Dodds et al. (P. Dodds & Danforth, 2009) analyzed tweet feeds to measure happiness based on the sentiment expressed in their tweets. Conover et al. (Conover et al., 2011) analyzed tweets for the purpose of predicting the political alignment of Twitter users. Pavlyshenko (Pavlyshenko, 2013) applied data mining methods to forecast events over Twitter. Ravikumar et al. (Ravikumar et al., 2013) analyzed tweet contents and users in order to come up with a ranking mechanism. Zang et al. (B. Zhang et al., 2013) presented a novel news ranking algorithm that utilizes tweets in order to come up with news article ranking.

### 1.2.3 Network Pattern Mining

For some of the most relevant articles: Yan et al. (Yan & Han, 2002) presented gSpan: graph-based substructure pattern mining algorithm, which is one of the most prominent pattern mining algorithms to date. Berlingerio et al. (Berlingerio et al., 2009) introduced a novel frequency-based patterns algorithm for mining graph evolution rules. This algorithm is capable of describing the evolution of large networks over time at a local level. Zhang et al. (J.-L. Zhang et al., 2010) presented FPG-Growth, a graph-based patterning mining algorithm to detect application level IO patterns. Our *HashnetMiner* in chapter 4 is based on the idea of contrasting two graphs, which was introduced by Ting and Baily in 2006 (Ting & Bailey, 2006).

## 1.3 Contributions

Our contributions in this dissertation are three folds:

1. We present a rule-based system that is concerned with digital recruitment on Twitter. The system solves a very important clinical problem of identifying subjects for participation in medical studies. Using live data streams, the system analyzes incoming tweets based on a given context (smoking cessation, diabetes subjects, etc). The system has at its core a knowledge acquisition engine which dynamically generates rules using data mining algorithms (e.g., association rule mining). The system is semantically enabled and can distinguish between (weed as a grass type and weed as a slang for marijuana) given the context. The system utilizes common hashtags to produces rules such as (e.g., (smoking and #weed), (smoking and #pot), and (smoking and #420). These rules are likely to be associated with (marijuana not cigarettes). This system is currently deployed to recruit/crowdsource individuals who are willing to participate in a study to help scientists understand the relationship between the environment and body weight.

2. We have designed a novel network model (we call it K-H networks) that is based on associations between hashtags and domain specific keywords that appear in tweets. The models aim to answer some of the following questions: Can we discover an unknown adverse effect of two drugs when taken together? Can we learn unreported drug side effects? The network is then used as a rich data model for exploration by means of pattern mining algorithms. We have also contributed a newly developed path mining algorithm, *HashnetMiner*, which operates on the network. The algorithm uses two different types of networks: (1) a keyword-keyword (K-K network) model which is used as a guiding model to highlight the positive controls. (2) The keyword-hashtag models (K-H network) compares certain patterns demonstrated on the K-K networks and extracts the negative controls (novelties). Such discoveries are the final output of *HashnetMiner*. We have used the algorithm to explore hidden paths in a drug-drug interaction network that is strictly constructed from tweets. The algorithm is designed to rank the discoveries based on their strength and return the Top-K patterns, where K is a configuration parameter specified by the end user.

## 1.4 Thesis Outline

The remainder of this thesis is organized as follows. Chapter 2 introduces a rule-based Twitter recruitment system for a nicotine patch study. This chapter discusses contemporary recruitment methods and their limitations. This chapter also discusses our approach for using Twitter to recruit people who are seeking professional help for their nicotine addictions. Chapter 2 presents the specifications of an expert system shell and the various algorithms used to accomplish this task. This chapter also presents our approach, which involves using association rule mining to generate rules from a large dataset of tweets to feed back into the system, causing it to be ever-evolving. Chapter 3 presents an exploratory analysis of the keyword-hashtag networks. This analysis is concerned with testing the degrees of separation between pairs of keywords in the network. Chapter 3 presents a newly developed algorithm, *HashnetMiner*, for mining keyword-hashtag networks. The chapter also

demonstrates how the algorithm is used to discover knowledge (drug interactions) from a large body of tweets. Chapter 5 presents the conclusions of all presented research aspects and future work.

## **Chapter 2**

# **A Rule-based Twitter Recruitment Intelligent System**

Digital recruitment is increasingly becoming a popular avenue for identifying human subjects for various studies. The process starts with an online ad that describes the task and explains expectations. As social media has exploded in popularity, efforts are being made to use social media advertisement for various recruitment purposes. There are, however, many unanswered questions about how best to do that.

In this chapter, we propose an innovative Twitter recruitment system for a smoking cessation nicotine patch study. The goals of the chapter are to (1) provide the system specification and design, (2) propose the approach we have taken to solve the problem of digital recruitment, (3) present two new algorithms, one for Twitter user ranking and the other for digital recruitment using social media, and (4) present the promising outcome of the initial version of the system and summarize the results. This is the first effort to introduce a practical solution for digital recruitment campaigns that is large-scale, inexpensive, efficient and reaches out to individuals in real-time as their needs are expressed.

A continuous update on how our system is performing, in real-time, can be viewed at <https://twitter.com/TobaccoQuit><https://twitter.com/TobaccoQuit>.

## **2.1 Introduction**

Digital recruitment is a popular online method that has been widely used for attracting individuals who are seeking products and services. Various online services have existed over the years to assist either individuals looking for jobs, or organizations seeking individuals who can offer services of interest. In recent years, digital recruitment is transforming from passive websites such as Monster, and CareerBuilder (CareerBuilder, 1995; Monster, 1999) to an active form as it is the case with LinkedIn (LinkedIn, 2003). Employers have direct access to individuals' profiles and can recruit those who meet their needs. More interestingly, with Twitter being built-in to LinkedIn, employers now tweet job-postings that they will be visible by hundreds of thousands people who can apply with their LinkedIn profiles in a matter of seconds.

Recruitment in the biomedical domain is still using more traditional methods of recruitment such as phone calls, emails and advertisements on Craigslist (Craigslist, 1995). There is room for biomedical research to use some of the online resources used by product vendors. This chapter is based on a specific project to improve subject recruitment for a smoking cessation study. Participants in the study will test two ways to use nicotine patches to help them quit. For recruitment we need to reach smokers who are interested in quitting, and then prompt them to seek additional information about the study. For similar studies in the past we have used newspapers, flyers, ads on buses, and Craigslist. We also hired a vendor to manage a GoogleAdWords campaign for us (Google, 1995). It is inexpensive to recruit with Craigslist, but its main disadvantage is that the results may have limited generalizability because it represents a special subsection of society – people searching Craigslist – and may not reflect society as a whole. Newspaper ads are of decreasing effectiveness with increasing costs every year. Google search ads managed by a vendor are effective for targeting

audiences, but costs are high (\$300-500/recruit). Flyers and bus ads are effective and of moderate to high cost (\$200-300/recruit).

With shrinking research funds available, we need to find an effective and inexpensive way to recruit study participants. An additional issue is that all recruitment efforts for human research must be approved by an oversight committee (an Investigational Review Board (IRB)). The use of social media for recruitment is new for board members, so they may be reluctant to approve new techniques to recruit with social media techniques. So far we have obtained permission to experiment with Twitter to identify smokers interested in quitting. During this testing, we sent public service messages on how to quit smoking instead of recruitment messages for the study. It is a challenge to create a program that won't be perceived by the board as coercive, offensive, or misleading.

**Previous Results** – As a yet-to-be published recent study, we did a combination of online and traditional recruitment and we received 2871 responses to the recruitment efforts; 1131 people were screened (39%) and 249 individuals (22%) were eligible. Of those eligible, 238 individuals (96%) participated in the study. We anticipate that the current study will be sufficiently similar to the previous study in that we will be able to make meaningful comparisons regarding contacts, screenings, and recruitment.

### **2.1.1 Recruitment Strategy and Specification**

We believe effective recruitment will require deeper understanding of the factors influencing a user response to an ad (Gupta et al., 2012). Generally speaking, for a given recruitment campaign this involves: (a) identifying interested users in the products or services, (b) catching the user in the time of need as they express their needs, (c) learning the user's take on the given recruitment message and incorporating this feedback, (d) increasing user's awareness of the products and services that fulfill their need, and (e) identifying large communities to target as a whole.

Using our innovative Twitter-based system, it is fairly easy to satisfy the criteria above and identify users who are expressing interest in quitting smoking. Twitter is a social media that supports the

existence of smaller communities, which we can discover computationally and target as a whole. Using the Twitter streaming API (Application Programmable Interfaces) we can identify those individuals who seek to quit in real-time. Learning the feedback from users about a tweet is retrieved in different forms on Twitter: (a) users can choose to follow an account or directly respond to a tweet positively or negatively using the reply feature or the messaging feature, (b) users can also favorite a tweet they view as it shows on their timeline, and (c) users can retweet a tweet and share it with their followers which is one of the most powerful features for not only expressing how much they like the Twitter campaign but also sharing the information with their circle of followers. Unlike Craigslist and other traditional recruitment platforms, Twitter users enjoy a sense of community. They can share ideas and trade experiences which could be very helpful in making a smoking cessation campaign successful. These expressions of ideas and experiences can flow over the network in a matter of seconds. Since our system is intended to interact with people, it is important that it does so in a more personal and human-like way.

### **2.1.2 Contributions**

The contributions of this chapter are as follows:

- We have designed a real-time smoking cessation recruitment system using Twitter to immediately fill-in users' needs. To the best of our knowledge, this is the first study that investigates smoking cessation recruitment using social media (e.g. Twitter).
- We present a non-textual, features-based scoring algorithm that computes the prestige score of a tweet based on Twitter's built-in features (Lists, Retweet, Trends, etc).
- We have incorporated domain expertise that engineers a rule-based system that decides which event is performed given an incoming tweet. This ensures that the system performs more intelligently in a human-like manner.



- We have demonstrated how we can derive rules from tweets that can make the system more intelligent

## 2.2 Recruiting Twitter Users

Before we explain our algorithm and approach, we next describe the problem setup in detail.

### 2.2.1 Problem Formulation

The focus of our study is to use Twitter to identify potential candidates, send recruitment messages, increase the awareness of smoking risks and compare our approach to other online traditional methods (e.g., flyers and newspaper ads). For each campaign, the goal is to identify and target Twitter users who are explicitly seeking to quit tobacco. In the following section, we formulate these tasks computationally. We use the following Twitter means to launch a campaign:

- **Explicit Tweet Contents:** We have inspected a large number of Tweets and manually selected very specific content to search for when a tweet is streamed.
- **User PageRank Within Twitter Graph:** Users who have a higher PageRank are likely to be influential ones.
- **Twitter Built-in Gears:** Retweets, Lists, and Trends are powerful means to discover communities of users. They present effective indications for how a given campaign is successful.

### 2.2.2 Proposed Approach

We designed a data stream processing software system that intercepts an incoming tweet in real-time to recruit users. The system then queries the tweet for some basic search keywords relevant to smoking cessation. There are various tasks the system must accomplish as each tweet arrives. The underlying algorithm determines which tasks are performed based on the exhibited context. Due to the daily limit of the number of tweets the system can send, the algorithm governs this limit

by maintaining a number of stochastic parameters. For each action the algorithm performs there is a specific parameter that dictates the rate of occurrence. However, it is unknown to the algorithm when a relevant event (Ritter et al., 2012) may occur and the rates could be too restricted and the system might not work effectively. On the other hand, if the rates are too loose, we reach the limit too quickly which will produce a “spammy” behavior that is not desirable. Currently, the algorithm supports tweets that are encoded in English. The following subsections discuss the expected input data, the output actions, and the building blocks.

**Expected Input Data** – The system operates on the following input data items and data structures

- Basic keyword list (e.g. smoking, tobacco, nicotine, quit)
- Explicit recruiting content list, which must contain a sentiment that expresses the wishes to quit smoking tobacco
- Pre-prepared database of tweets to circulate
- Real-time streaming tweet
- List of trending topics

**Expected Output** – The system produces tweets, retweets and replies in real-time to users who are identified to be strong candidates.

### **Twitter Logistics**

Twitter has features, rules and regulations to ensure a good experience for its users. These rules must be observed for the algorithm to work correctly. We discuss the various features that each user can perform, and how the algorithm uses them in each computational step.

- **Tweet** – This is the feature that makes Twitter the prominent social media platform it is today. A microblog of 140 characters allows the user to voice their opinion to the Twittersphere.

When a user sends a tweet out, the tweet is displayed on the user's timeline and it is viewable by all of his/her followers. The tweet may also be viewed by any other Twitter user searching for keywords that match some content in the tweet.

- **Reply** – When a user tweets an update, followers can directly respond to that tweet. For the reply to reach the receiver, the original user's screen-name is appended to the tweet and preceded by the at (i.e. @) sign (e.g. @MyTwitterScreenname).
- **Mention** – Similar to the Reply feature, any Twitter user can share any content with a specific Twitter user by simply mentioning their screen-name anywhere and it does not have to be in response to a Tweet.
- **Retweet** – This feature is one of the most innovative features a social media platform has ever invented. When a user receives a tweet that might contain valuable information, they have a way to share it with an entire network of followers. Some tweets get retweeted thousands of times.
- **Lists** – Twitter enables its users to group other users who share a similar interest.
- **Hashtags** – Twitter treats any keyword(s) preceded by the (#) as a special string for annotating tweets and making it easy to search, track and follow. Such special keywords are called hashtags.
- **Trends** – Twitter shows hashtags, words, names of people, or any topic that is trending in real-time. The trending list is always updated as new hashtags start to trend and older ones die out.

Additionally, Twitter has certain rules and restrictions that must be observed by its users. This is crucial for any algorithm that attempts to automate sending tweets, retweets or perform any other actions using the Twitter API. The following are some of the most important ones that our recruiting algorithm observes.

- **Max Daily Limit** – Twitter restricts the number of tweets that can be sent from one account to 1000/day.
- **Duplicate Status** – Twitter does not allow tweet duplicates within a given duration of time. Each tweet must be unique in all or partial content for Twitter to allow it to be posted.

### 2.2.3 System Architecture

This section describes the architecture of our system and illustrates how online recruitment can be done in a time-aware fashion. There are three different steps: **Twitter Monitors** is a software component that keeps track of the tweet streams, Lists, and Trending events and words. **Tweet Analyzer** is a filtering component that queries the real-time tweets for specific phrases that explicitly indicate calls for help to quit smoking. **Event Processor** is the *online* transactional component that sends recruitment messages to Twitter users. Next is an elaboration on each component individually.

#### **Twitter Monitors**

The system we designed is real-time recruitment software that reaches out to those Twitter users who are soliciting advice, help, or products to give up smoking. Therefore, the Twitter monitors are designed to read the streaming tweets in real-time. Since the system is concerned with smoking cessation, we designed the monitors to track those tweets that have related keywords (smoking, tobacco, quitting, addiction, cigarettes, etc). The system also monitors real-time trends (X. Yang et al., 2012) to keep track of emerging events and news that are encoded in the English language and occurring within the US. Using the publicly available Twitter REST web services (Twitter.com, 2006), and Twitter4J Java wrapper API's (Yamamoto, 2007), we developed the Twitter monitors.

#### **Tweet Analyzer**

Due to the massive amount of tweets received by the monitors, further analysis must be performed to filter out the irrelevant tweets. Once captured, the analyzer then groups the tweets into three groups:

1. **Platinum Tweets** – those tweets that contain contents that solicit explicit help to quit (e.g., “I must quit smoking tobacco now”).
2. **Golden Tweets** – which are the tweets that contain contents that indirectly solicit help to quit (e.g., “smoking makes me cough my lungs out”).
3. **Info Tweets** – all other tweets that contain useful information which can be shared with followers.

The Tweet Analyzer is a simplified version of the *Text Classifier* software and is based on a regular-expression dictionary look technique. We developed a home-grown component using Java Regular Expression and String manipulation. Upon the completion of this step, the tweet is passed to the event processor component along with a label (e.g., *token*).

#### **Database of Tweets, Hashtags, Web Links**

As mentioned above, the proposed recruitment system operates mainly on streaming tweets. Nevertheless, it needs a backend database to be fully automated. The team has created a large number of pre-prepared tweets to research target Twitter users. One type of this interaction is to perform soft-recruiting by periodically announcing our recruitment services to followers and the outside world. These tweets contain the following contents (1) uplifting informal messages to share, (2) a 24/7 voice service number for people to call, (3) a tiny URL that links users to the website of the services, and (4) several hashtags that we selected from existing hashtags as well as some of our own. The hashtags are used based on the context, day of the week, and ongoing events. (E.g., #Lent, #TGIF #FollowFriday to use on Friday, and #HappyMonday to use on Monday) are appended to the tweet body dynamically. In addition, there is a new type we call direct recruitment for the tweets that are sent to a Twitter user in response to a tweet that seeks help. Another type of tweet is one that enables democratization of knowledge and shares the most important news, research studies and other services to increase the awareness of tobacco smoking risks. All three types of tweets are

generated on the fly using the links, pre-prepared tweets and hashtags. Table 2.1 shows some sample recruitment messages we send in different contexts.

<b>Tweet Type</b>	<b>Tweet Text</b>
User Tweeting	I want to stop smoking again.
User Tweeting	Why is it so hard to quit smoking cigarettes :( Someone help me ...: 17 hours ago ... I have good reasons to quit.
User Tweeting	I should give up smoking cigs for lent .... SIKE !
Soft Recruitment	Make friends with your lungs; quit smoking!
Soft Recruitment	Step out of the past and into a smoke-free future.
Soft Recruitment	Breathe easier. Quit smoking!
Direct Recruitment	You can quit now. Be smoke free quickly. call #877-437-6055 #smokefree
Direct Recruitment	!!!!Good idea! It's time to quit. Here's help: bit.ly/11Z2GAY #lent
Direct Recruitment	Take action to help a loved one quit: bit.ly/11Z2GAY#socialcare

Table 2.1: Pre-prepared Recruitment Sample Tweets

## Event Processor

When a tweet is classified to be Platinum, Gold or Info, a token is sent to the Event Processor component to perform an action. Based on what the token is, an action is performed. This component is powered by an algorithm that performs a sequence of tasks until the event is processed. Starting with the token received, the algorithm activates an action and dynamically binds the action with database contents and checks the current rate for this particular type of event. It is common to mine patterns in data streams based on streaming textual feature selection methods (Yu et al., 2012). We follow a similar approach here but we look for non-textual features (e.g., user number of followers, number of friends, on a specific Twitter List). Based on the existing features associated with each tweet, the tweet gets a final score. Additionally, the processor crawls any tiny URLs that exist in the received tweet body to compute a local approximation of PageRank to incorporate into the prestige of the tweet. For this task, we have utilized an open-source web crawler Java library called

Crawler4J (Ganjisaffar, 2012). Figure 2.1 shows the architecture and the various components. The algorithms are explained in the upcoming sections.

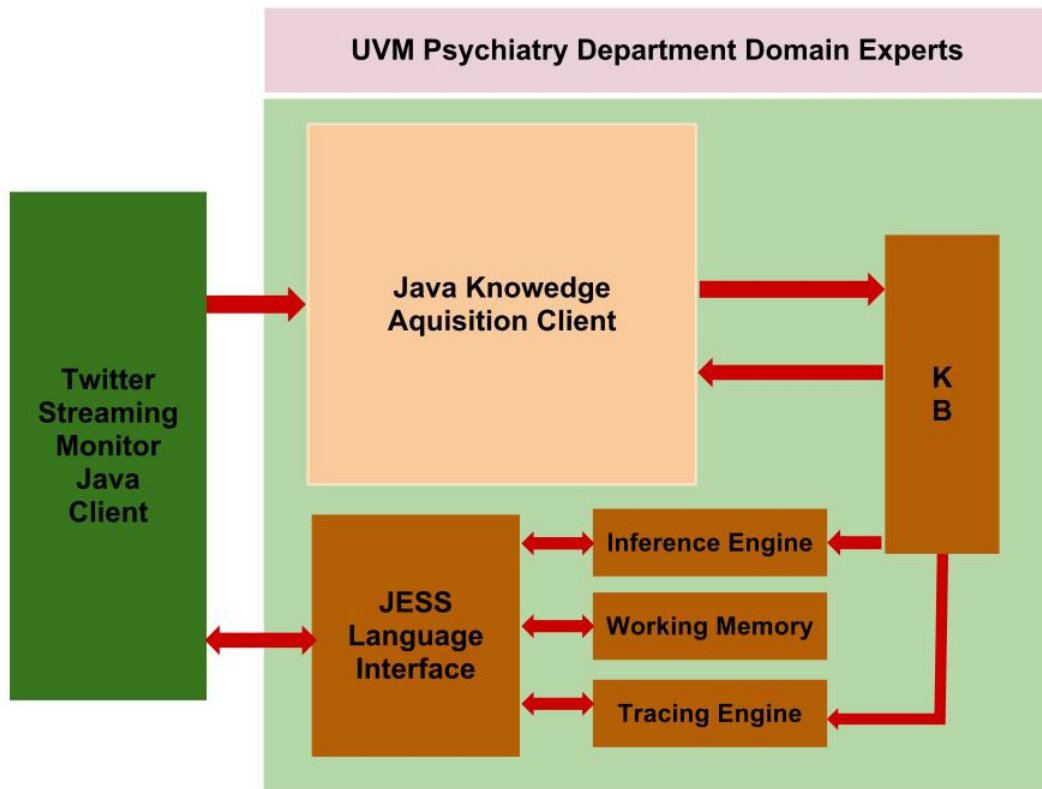


Figure 2.1: Twitter Recruitment System Architecture

## 2.3 System Performance And Evaluation

The system is currently running on a test environment at <https://twitter.com/TobaccoQuit>. This section gives an account of how the system is performing. We have tweeted about 23K tweets, some of which are our own messages from the database, replies and mentions. We have about 1100+ followers, while the average Twitter account has 126 based on the Twitter statistics of 2012 (Pring, 2012). In a period





maintaining were direct messages to ask questions privately. Most of the impressions received were positive. We also monitored the clicks to the service URL link using <https://bitly.com/ZDMaA7+/globalBitly>\* (Incorporation, 2013). We launched our campaign in May, 2013 and we have received around a 1000 clicks on the study's website. We found that 81% of traffic was generated by Twitter by itself. The remaining 19% of the traffic was generated out of email clients, mobile devices, Instant Messages and other. Figure 2.3 shows the daily numbers of clicks for the traffic ratio.

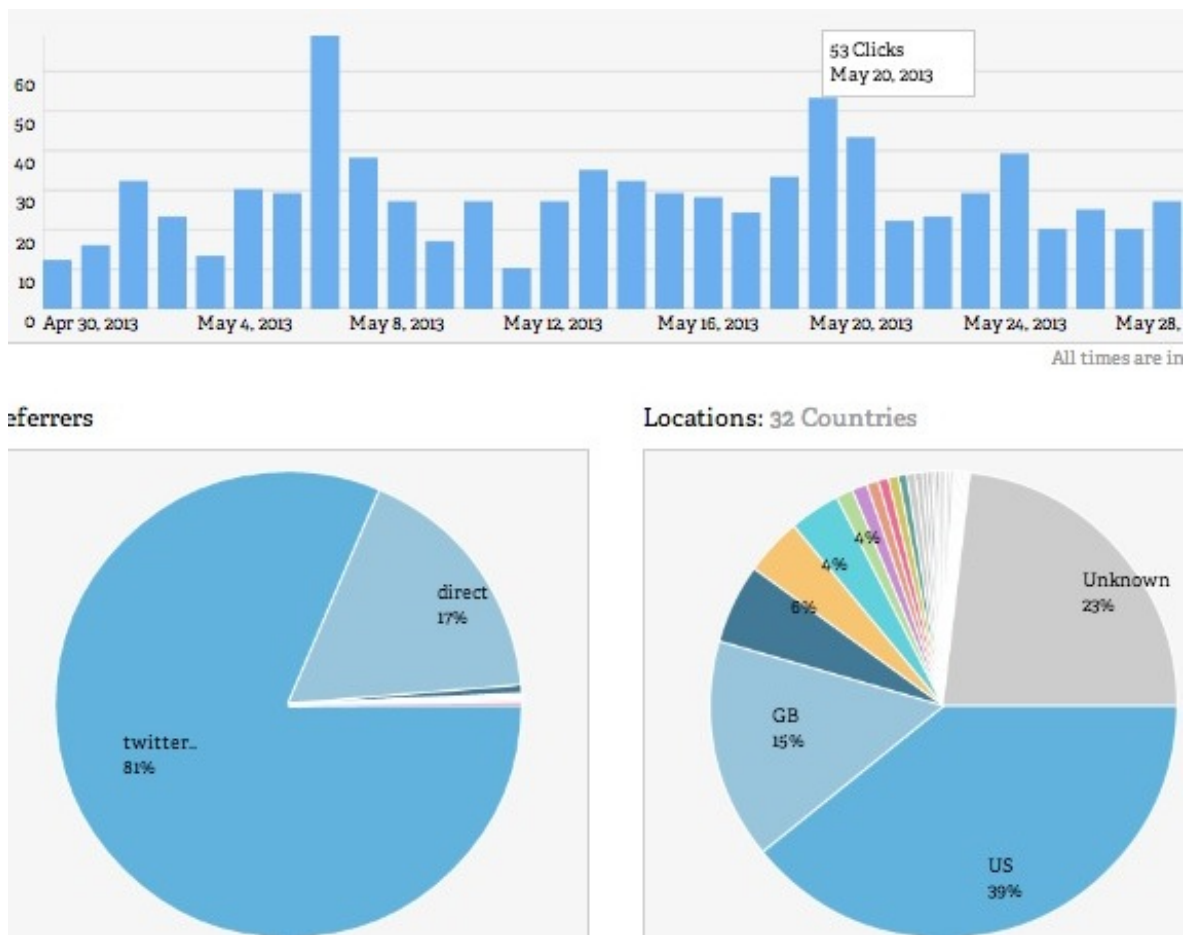


Figure 2.3: TobaccoQuit Traffic using BitLy

\*Real time update: <https://bitly.com/ZDMaA7+/global>

## **2.4 Twitter Recruitment Algorithm**

The purpose of this algorithm is to exhibit intelligent behavior that is considered acceptable by the general public (Twitter users). The algorithm should not propagate spam, follow or recruit irrelevant users. It also must obey Twitter rules in order for the Twitter account to remain lively and influential. More importantly, it must demonstrate courteous behavior to its followers and to the users who are seeking help to give up smoking. For example, if a recruitment message is sent immediately after a tweet that is identified to be relevant, it elicits a negative reaction from users. Therefore, a delay function is used to relax the behavior of our system when it tweets from the @TobaccoQuit Twitter account. Another important goal of the algorithm is to enable the sharing of knowledge by sharing tweets that may contain important information on the risks of lung cancer, new research studies on smoking, and uplifting experiences shared by former smokers.

When a streaming tweet is intercepted, it is immediately searched for the basic set of keywords and the language that the tweet is identified. If the algorithm classifies a tweet to be relevant, it computes its prestige using the scoring algorithm that is described next. Based on this score, the algorithm performs a Retweet (RT), a Mention, or a simple Tweet action as a quitting tip to the followers. If it is indeed the case that it is a Mention task to be performed, an extra step is needed to further classify the user to decide whether he/she should be recruited. The backbone of the algorithm is a rule-based system that can decide which action to perform based on the context. Next is the pseudocode and the flow of Algorithm 1.

### **2.4.1 Non-Textual Feature Scoring Algorithm**

Various studies have attempted to measure the prestige of a tweet. Some studies assumed that if a tweet has a URL then it is important. Although this might be true to some extent, further analysis is necessary. Other studies considered the number Retweets to a tweet as signifying the importance of a tweet. This is indeed interesting but an agent program can simulate this behavior and circulate

---

**Algorithm 1** Data Stream Recruiting Algorithm

---

**Input:**

$W$ : basic keyword list

$A$ : action {Direct, RT, Soft}

$C$ : user\_class { $y, n$ }

$r$ : relevance threshold

**Description:**

```
1: Foreach tweet  $t$ 
2:   search( $t, W$ )
3:   If relevant
4:     score = prestige( $t$ )
5:     classify(user)
6:     If class( $y$ )
7:       perform_action(Direct)
8:     Else If score >  $r$ 
9:       perform_action(RT)
10:    Else
11:      perform_action(Soft)
12:    End If
13:  End If
14: End Foreach
```

---

spam using RT. However, the above methods have inspired Yang et al (M. Yang et al., 2012) and Yamaguchi et al (Yamaguchi et al., 2010) to recycle the idea of using the RT feature as an indication of a high prestige.

We propose the following algorithm to estimate the prestige of an incoming tweet. The algorithm takes as input two different parameters: (1) the user in-degree, which is essentially computed from network of followers, and (2) the tweet body which includes all the words, hashtags, and links embedded. The algorithm uses two different lookup lists: a list of trusted users, and a list of irrelevant users. When the algorithm runs, the lists are initialized and populated by the corresponding users. The algorithm returns the final prestige score for the given tweet and hence the Retweet function decides whether to RT it or not. The following is the theoretical background to show that the in-degree is a good approximation of the global PageRank.

## PageRank Theoretical Background

**PageRank score** – This score is determined by summing up the PageRank scores of all pages that point to  $i$  (Wu et al., 2007; Wu & Kumar, 2009). The score of page  $i$  [denoted by  $P(i)$ ] is defined by:

$$P(i) = \sum_{(j,i) \in E} \frac{P(j)}{O_j}$$

s.t.  $O_j$  is the number of out-links of page  $j$

**In-degree PageRank score** – The average PageRank of a page with in-degree  $k_{in}$  can be well approximated (Fortunato et al., 2005, 2007) by the following closed formula:

$$p(k) = \frac{q}{N} + \frac{1-q}{N} \frac{k_{in}}{\langle k_{in} \rangle}$$

s.t.  $N$  is the total number of pages,  $1 - q$  is the so-called damping factor, and  $q$  is the probability of jumping from a node to another in a random walk.

**Graph Model** – We treat each tweet as its own local directed graph DAG as follows:  $\forall$  Twitter user  $u_i$ ,  $\exists U = \{u_i, \dots, u_k\}$  such that  $U$  is the set of  $k$  nodes. The relationship between  $u$  and  $U$  is established by Twitter's *Following* relationship. By applying the in-degree approximation formula of the PageRank algorithm, we get the score of user prestige in the Twitter network.

**Boosting, Penalizing, and Rewarding** – The algorithm boosts the score of a user who happens to be in our trust list. If the user is tweeting trending content, then the user score gets rewarded for such content. Finally, if the user happens to be on the irrelevant list, the score is severely penalized by 0.25. The overall score return is 1. It is important to note that the lists are designed to be mutually exclusive. If a user exists on a trust-list, he/she is not on the lists and vice-versa. The pseudocode is presented in Algorithm 2.

---

**Algorithm 2** Tweet Prestige Scoring Algorithm.

---

**Input:***T*: trusted user list*N*: negative user list**Output:**

prestige\_score

**Description:**

```
1: Foreach tweet t
2:   u ← get_user(t)
3:   If u ∉ N
4:     score ← score + 0.10
5:   Else If verified_profile(u)
6:     score ← score + 0.10
7:   Else If u ∈ T
8:     score ← score + 0.10;
9:   Else
10:    score ← score - 0.25
11:  End If
12: End Foreach tweet t
13: PR ← compute_PageRank(u)
14: prestige_score ← score * 0.25 + PR * 0.75
15: Return prestige_score
```

---

### 2.4.2 Rule-based System

The heart of the recruitment system lies a rule-based component that communicates with the textual phrases in the tweet body. When a tweet is analyzed by the Tweet Analyzer component, the output is sent to the rule-based system to decide which task should be performed. When this is completed, a decision is made and a token is sent back to the Event Processor. This system is comprised of 5 main rules that correspond directly to the Twitter features that the recruitment system is based on. We implemented this system using JESS, a Java Expert System Shell (Hill, 2003). Due to the page limitation, we only discuss the Tweet, Reply, and Retweet rules.

## **Tweet Rule**

This is a direct implementation of the Tweet feature to soft-recruit individuals without being “spammy”. Since it is able to read and analyze the streaming tweet content in real time, the tweet rule is able to decide when to tweet a message to the world if someone is seeking help quitting tobacco. Although the tweets sent out are not a direct reply to the user seeking quitting information, it contains Twitter hashtags that will match the sender’s hashtags. This will increase the chances of finding our tweets while this particular user is still active on Twitter.

---

### **Algorithm 3** Tweet Rule Implementation.

---

```
(defrule action-soft-recruit
  (twitter-user (language "en")
    (screen-name ?screenName))
  (raw-tweet-info
    (id ?tweetID)
    (text ?tweetText))
  (test (and (not
    (contains-retweet-keywords
      ?tweetText))
    (not
      (contains-article ?tweetText))))
  =>
  (assert (recruitment-action
    (action "soft-recruit")
    (tweetID ?tweetID))))
```

---

## **Reply Rule**

Currently, the reply feature looks at simple specific tweets that contain messages from users who are explicitly soliciting help to quit tobacco. Once identified, the rule is activated and the action is triggered as “Reply”. All facts in the working-memory of the knowledge-based system are bound to this Reply only. This makes composing the reply to the tweet possible. The rule makes use of another rule which decides if a tweet has the basic and specific contents. The reply rule must match this pattern to perform basic filtering before the action is activated.

---

**Algorithm 4 Reply Rule Implementation.**

---

```
(defrule reply-action
(golden-tweet-info
  (id ?tweetID)
  (text ?tweetText))
  (twitter-user
    (language "en")
    (screen-name ?screenName))
=>
(assert
  (recruitment-action
    (action "mention")
    (tweetID ?tweetID))))
```

---

**Retweet Rule**

Twitter's Retweet feature (aka RT) is one of the most influential features of Twitter for getting the word-of-mouth circulated quickly to so many users. RT is a very powerful tool such that when a user searches for a tweet or receives one, they can choose to share it with their followers via this feature. Once a tweet is retweeted it shows up to all the original sender's followers. After identifying a relevant tweet, this rule further decides whether it should be shared in the form of RT or not.

---

**Algorithm 5 RT Rule Implementation.**

---

```
(defrule action-retweet
  (twitter-user
    (language "en")
    (screen-name ?screenName))
  (raw-tweet-info (id ?tweetID)
    (text ?tweetText))
  (test (and (contains-retweet-keywords
    ?tweetText)
    (contains-article ?tweetText)))
=>
(assert
  (recruitment-action
    (action "retweet")
    (tweetID ?tweetID) )))
```

---

## **2.5 Association Rule Mining Knowledge Acquisition Using Apriori Algorithm**

Twitter is a dynamic social media where its contents are constantly evolving. Having a fixed set of rules deliver by domain experts will not suit this ever-evolving nature. Rules must be acquired from the tweets themselves in addition to the knowledge by the experts. We performed a series of rule mining experiments to come up with the rules we add to the knowledge base. It is essential to note that tweet contain various types of contents some of which is plain text, URL links, and special type of words are called hashtags. A hashtag is a word or a collections of words proceeded by the symbol (#) and concatenated together with underscores or using a upper-lower case convention. Hashtags seem to carrying various semantics and we believe they are worthwhile exploring such components in order to generate rich rules. The next sections explain how to we use experiment with hashtags and how to evaluate valuable rules using Apriori, and Association Rule Mining algorithm.

### **2.5.1 Data Description and Gathering**

Collected a two sets of streaming tweets captured by Twitter’s streaming API. First set is gathered using ten search terms: (quit, quitting, smoking, nicotine, patches, smoke, cigarette, cig, cigs, ecig, marijuana). Second set of tweets is a superset of the first one and is comprised of 30 terms which include the terms from the first set. Given those set of keywords, it is possible to get tweets that contain hashtags that are identical to the search keywords (e.g., patches and #patches). It is also possible to see entirely different hashtags that may not have any syntactic similarity with the input search keywords (e.g., smoking and #weed). Our method distinguish between these two types of hashtags to see if one is better than the other and quantify that. According to our method of analysis, a tweet such as (“you can stop up smoking now, free #patches call #800quitnow”) generates the following records for the given keywords we used: (1) (smoking, #patches), (2) (smoking, #800quitnow). We treat such instances as association transactions. When a large set or tweets are analyzed, they also



generate a large set of transactions which we analyze using Apriori Algorithm (Agrawal & Srikant, 1994). We collected two data sets from a 10,000, 25,000 tweets. Records resulted from each type is formatted using Weka's ARFF format to analyze using the Apriori algorithm. The output of some of experiments is shown in Figure 4.3

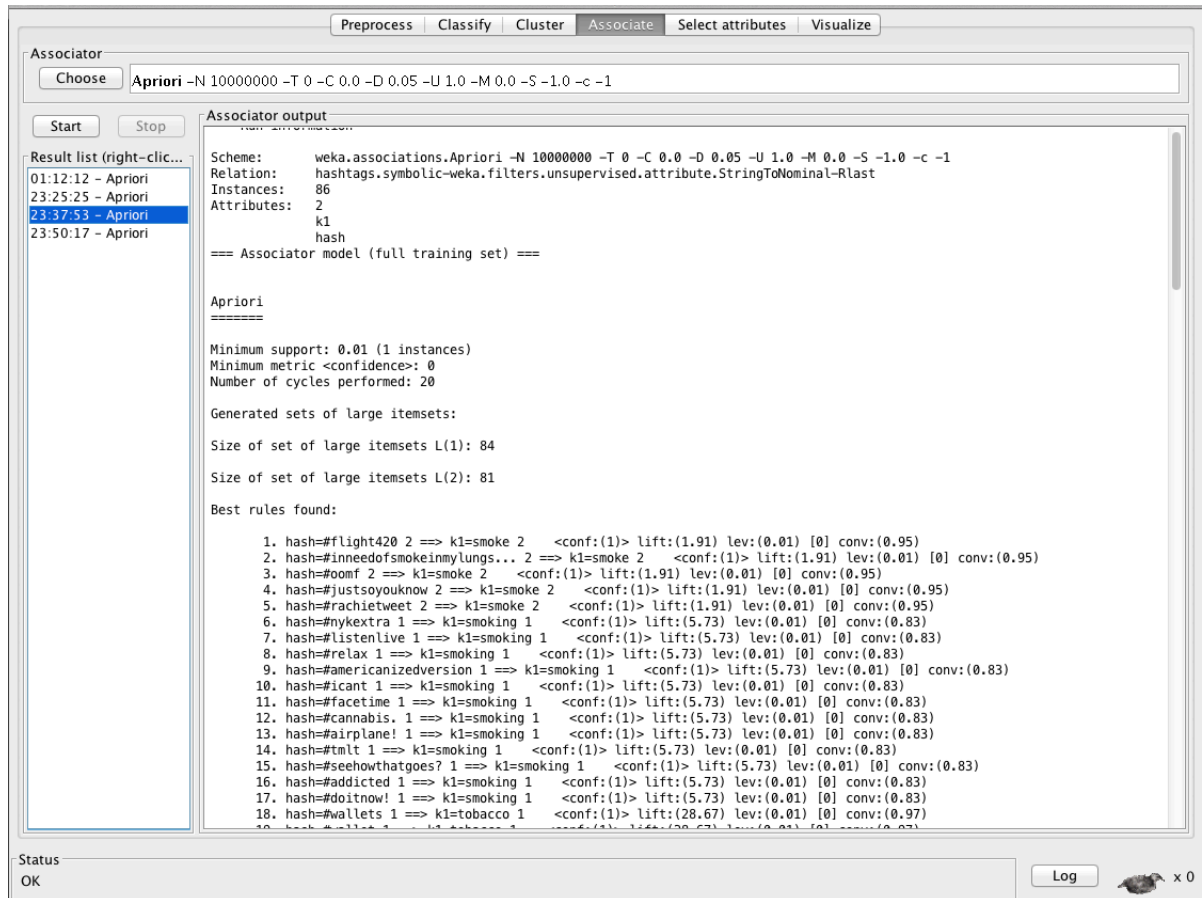


Figure 2.4: Association Rule Mining using Apriori in Weka

## 2.5.2 Setting Up Minimum Support

Finding frequent item sets in tweets is a challenging task. This is due to the fact that tweets encompass a large spectrum of topics. Each topic is covered by a wide range of users from different cultures, backgrounds. Clearly this is different from the fixed number items people can buy from

any grocery store or online on amazon.com. We found that a very small fraction of minimum support is sufficient to expose the local trends about the topics that we are interested in (i.e., smoking cessation in this case). We have experimented with the following minimum support values (0.01, 0.01, 0.0001). This is sufficient to expose interesting patterns and associations found in the dataset analyzed. Table 2.3 shows a comprehensive list of parameter values. We performed 3 consecutive experiments on each data set by fixing the minimum-support parameter and continuously dropped the minimum confidence with a small fraction to measure the performance of each type. Table 2.3 shows the various settings of minimum support and confidence of the experiments.

<b>Min Supp</b>	<b>Min Conf</b>														
0.01	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.01	0.001	0.0001	
0.001	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.01	0.001	0.0001	
0.0001	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.01	0.001	0.0001	

Table 2.3: Experiments settings

### 2.5.3 Experimental Results

Figures 2.5 and 2.6 show the comparisons of confidence using difference support and confidence levels. The curves on the left show both the average confidences of each experiment. The horizontal axis reflects the a single parameter setting of an experiment and vertical reflects the average confidence of all rules learned form this experiment. The curve in blue shows the average confidences for the associations of keywords and their identical counterpart hashtags. The curve in red shows the average associations for keywords and the entirely new hashtags. The histograms in blue reflect confidence gain for data set 1 and the histograms in green reflect the confidence gain in the large data set.

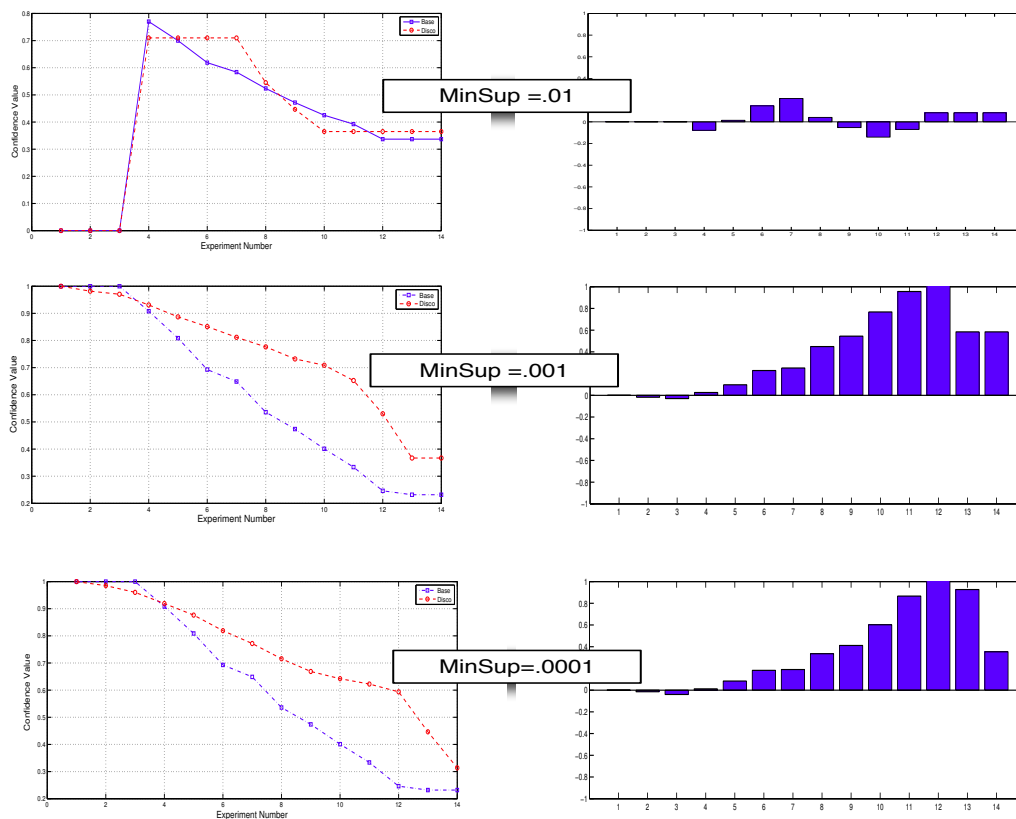


Figure 2.5: KK vs KH dataset 1 experiments with three minsup levels showing the gain ratio favoring KH

## 2.6 Discussions

We have presented an emerging smoking cessation application that is based on Twitter. We are taking an incremental development approach and deploying it in a test environment to obtain the general public's feedback on the services we launch. This approach is intentional, as we must comply to the recruitment protocol proposed in the NIH proposal specification and the IRB board. We proposed two novel algorithms to rank social media users and to develop recruitment services of any type. Our smoking cessation recruitment system can also be extended to launch services for treatment of drug and alcohol problems and other medical recruitment services.

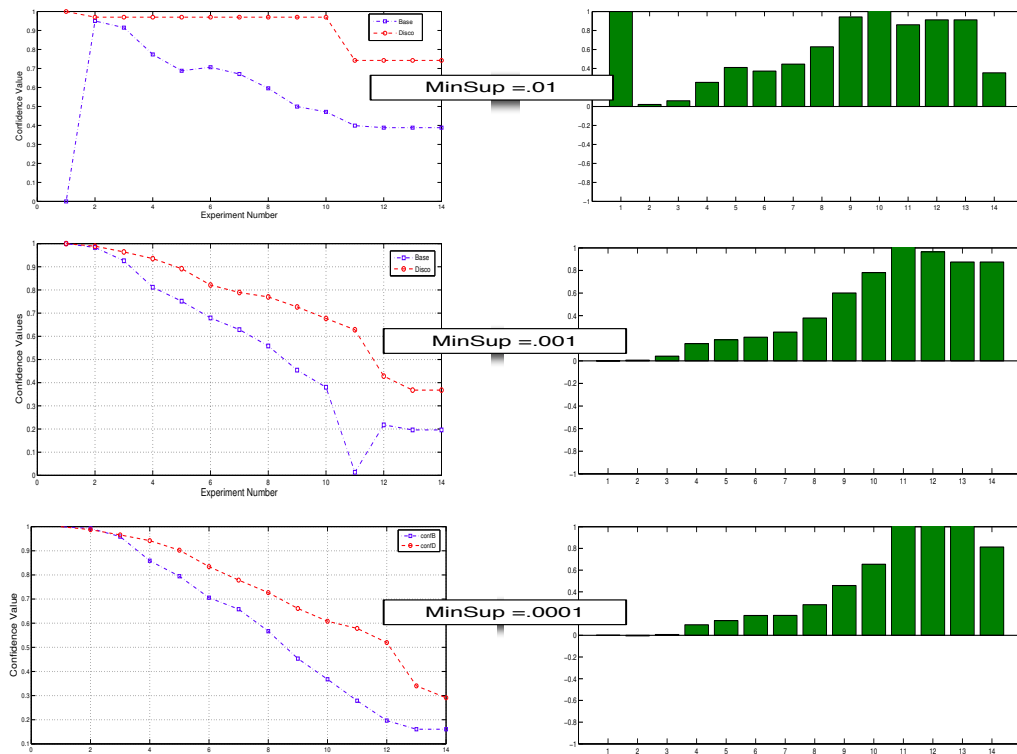


Figure 2.6: KK vs KH dataset 2 experiments with three minsup levels showing the gain ratio favoring KH

Twitter does not expose the impressions API to the public yet. Therefore, the current software still requires minimal manual labor. We have a designated person that carefully reviews the various impressions and manually responds to general Twitter users or followers. Another limitation is that the system is unable to share external web news beyond the Twitter platform. However, we get around this by using Google Alerts, which delivers new web articles to a designated mailbox. News web links are manually annotated and deposited into the database to be shared.

We have not acquired access to a web link database. This limitation prevents us from computing the in-degree PageRank for links contained in the tweet body. We are resorting to compute pseudo-rank based on the out-degree links. Another limitation to the current algorithms is the lack

of adaptability in response to tweet traffic. This would help because the tweet rate varies during the day. Fixed rates could be high during a high traffic time, which would be viewed as spammy. At other times, the services might not be effective if the rates are too low when traffic is low. We hope to address this as we learn our capacity to handle recruitment requests from Twitter users.

From a data mining and knowledge acquisition perspective, the experiments performed on two sets of experiments have not shown a consistent trend especially the ones with lower support (0.01). In both types of experiments we observed fluctuation in the average confidences. Some fluctuation was for the rules that contained identical hashtags others were for the experiments that contained entirely new hashtags. We also believe that the average confidence abstract much of rules specificity which stands in the way of learning which rules are useful for the system. The average measure fails to highlight which rules are significant than others. In order for the system to mature we must know which rules are significant which could not be achieved using the average confidence measure. Further experiments and measures will be explored in future studies to incorporate hashtags and also computation evaluate the ones produced by the association rule mining experiments such as the ones above. Manual inspections of the the association rules that were generated have been found more interesting and worthwhile stating here: (1) The term Marijuana that was highly associated with the following hashtags (#weed, #pot, #grass). Most Twitter users refer to Marijuana but one of those words for a slang. Marijuana was also found associated with the hashtag (#cannabis) which is the plant organism that people use to smoke and inhale the marijuana substances. More interestingly, marijuana was also found highly associated with a hashtags called (#420), after investigating the hashtag, we found out that this was a reference to the 20th April, which is a code-term used primarily in North America that refers to the consumption of marijuana and by extension, as way to identify oneself with cannabis subculture.

## **2.7 Acknowledgments**

This work is funded by National Institute of Health (NIH) Grant: 1 R01 CA165080. We also thank the HPBL team for populating our database of tweets. Special thanks to the Twitter4J project lead, Yusuke Yamamoto, and all contributors.

## Chapter 3

# An Exploratory Analysis of K-H Networks

Social Media Big Data have transformed the scale of exploratory analysis on the web and offered new means of performing tasks that were not feasible before. Over a half century ago, Milgram showed that the average number of intermediaries (called the “degrees of separation”) between two individuals is less than six (between 4.4 and 5.7). The study was done via means of regular mail on a very limited number of participants. With the advances of Internet, Dodds et al. performed the first electronic Milgram’s using 60,000 participants and achieved similar results of Milgram’s. These experiments remained within a very small-scale until the emergence of Facebook and social media. In 2012, Backstrom et al. carried out the largest Milgram-like experiment performed using the entire Facebook network. The findings showed that the average degrees of separation is 3.74, which suggest a shrinking world due to the impact of Big Data in a highly connected environment. Inspired by this recent finding, we pose the question: How does this newly observed phenomenon affect contents on Big Data environments? In this study we present an exploratory analysis of large-scale K-H networks gener-

ated from Twitter. We used two different measures (1) The number of vertices that connect any two keywords, (2) The eccentricity of keyword vertices, a well known path measure. Our analysis shows that K-H networks conform to the phenomenon of the shrinking world. Specifically, it shows that the number of vertices of any two keywords, that were not originally connected in the K-K networks, is exactly three while the eccentricity of every keyword is exactly four.

### **3.1 Motivation**

In recent years, hashtags generated by Big Data environments (Twitter, Facebook, and others) have flooded the web. Hashtags have become an essential feature of contents generated by most social media platform. Particularly, people have used hashtags to express their interests, feelings, and experiences about virtually any topic. Just to name a few, people used hashtags to announce the cars they drive (e.g., #honda), the movies they watch (e.g., #TheHobbit), the hobbies they enjoy (e.g., #skating), and the foods they cook (e.g., #fish, #grilling). As a result of such social tagging, hashtags have attracted the attention of many scientists in various domains (Conover et al., 2011; Kleinberg, 1999; Keyhole.co, 2007; Carter et al., 2011). Hashtags are not random. They are related to the tweets in context. Therefore, it is difficult to study hashtags without such context in mind. This point inspired us to study hashtags in relation to the words they appear with. Hashtags, however, are quite arbitrary and cannot be predicted. Thus, specifying the context in advance and capturing the hashtags in the context is a desirable approach.

One may study hashtags associated with any subject of any domain. The subject may range from medicine, life science, pharmaceutical, products, services, accessories, etc., virtually anything as long as one has a list of terms (or words) to fetch the tweets in the context. This list of terms can be from any structured list of vocabularies, online ontology (e.g., BioPortal), a database of products or services (e.g., Honda, iPhone, over-the-counter drugs, books, movies, music), and many more. This approach can generate many associations of the words of interest and the emerging hashtags. Word



association networks are a common formal model to study the association between words (Priss & Old, 2007; Deyne & Storms, 2008; Palla et al., 2007). A hashtag is not particularly a word but is rather a symbol which may refer to a word. In order for us to understand the difference and gain some insights, we have constructed two different types of association networks: (1) a network of association between words that belong to the context of study and, must also appear in the same tweet to count as association and (2) a network of association between word and hashtag both of which also must appear in the same tweet to count as association.

While conducting our exploratory analysis of the association networks of words and hashtags, we came across a very interesting article presented at the Web Science Conference, 2012 by Backstrom et al. with the title: “Four Degree of Separation” (Backstrom et al., 2012). They carried out the largest Milgram-like experiment ever performed using a Big Data environment. Their study analyzed the entire Facebook graph in regard to the small-world phenomenon and reported that the average number of intermediaries (called the “degrees of separation”) between any two individuals is 3.74. This particular finding was exciting to us since it was performed on a Big Data platform that is comparable with Twitter in many ways. The finding also motivated us to ask the following questions: (1) What does the shrinking-world phenomenon mean in terms of words associations? (2) What kind of impact does it have? If the world is indeed shrinking, does this mean that the world of words is also shrinking? (3) Do hashtags also contribute to the acceleration rate of this shrinking world phenomenon? Satisfying these curiosities empirically may lead to providing interesting insights, answers strictly from tweets to such questions as: “Can we find the best hashtags for a given topic to maximize the viability of a tweet?”, “Can we produce the base components for new recipes?”, “Can we learn the adverse effects of two drugs which may interact when taken together?”, and many more.

In this chapter we report our findings in studying the two types of networks: (1) keyword-keyword (K-K) association networks and (2) keyword-hashtag (K-H) association networks. The K-K networks are constructed from plain words that co-occur in the same tweets and turn out to

be associated using the association analysis Apriori algorithm. The K-H association networks are also generated from the same set of tweets using the same experiment configurations, but they include the hashtags that were found in the tweets. In view of the shrinking-world phenomenon, the primary focus of our findings is on the the number of vertices that makes up a path of any two keywords. In particular: for any pair of keywords (i.e., their concepts) disconnected in a K-K network, if they are connected (i.e., there exists a path between them) in the K-H Network, how many vertices that make up this path as a lower bound measure? Additionally, we are also interested in the eccentricity (Gross & Yellen, 2005) of vertexes – which we use as a conservative measure indicating the maximum shortest distance of a path that starts with a keyword as an upper bound. From these two observations we reach a conclusion that if there exists a new path between any two keywords in the K-H network, the number of vertices that make up such a path is exactly three, while the eccentricity of each keyword vertex was exactly four. This is indeed a smaller degree of separation than reported by Backstrom et al (i.e., only 2 links away from the source keyword of a path of three vertices).

## **3.2 Related Work**

Small-world networks have generated a great deal of interest among scientists of different domains (e.g., mathematics, social science, chemistry, physics, biology) (Strogatz, 2001; Newman, 2003; Marvel et al., 2013; Newman, 2000; Kleinberg, 1999; Perakis et al., 2014). These networks exhibit unique properties such as robustness (i.e., high clustering coefficients) as well as efficient transport properties (i.e., low average path length). This massive wave of interests was originally sparked by the six degrees of separation experiment by Travers and Milgram (Travers & Milgram, 1969). Their experiment on the small-world phenomenon is most prominent background for our work in this chapter . In the experiment he aimed to deliver a letter in the regular mail to a certain person. The experiment was restricting people from sending the letter to the destination directly unless the individual sending the letter knew the target person at the final destination in person. The experiment

results showed that the average path length is roughly six from the first person sending the letter until it reaches the target person. This surprising finding has found a large appeal among scientists who are interested in networks and their topologies. Accordingly, there have been some efforts to reproduce Traverse and Milgrams' work in various ways (P. S. Dodds et al., 2003; Backstrom et al., 2012). Dodds et al. (P. S. Dodds et al., 2003) performed a similar social-search experiment in which more than 60,000 users aimed to reach one of 18 target individuals in 13 countries by forwarding messages to acquaintances by means of electronic email. The experiments showed that an email message can reach the target in the median of five to seven steps.

With the widespread use of social-media and the accessibility to large-scale social network data, the previous work has inspired people to do similar experiments on much larger scale, where a network is comprised of millions of nodes. Particularly in 2012, at the WebScience Conference in Chicago, Backstrom et al. (Backstrom et al., 2012) presented their exciting work on the largest Traverse and Milgram-like experiment ever performed. The authors reported the results of the first world-scale social network in terms of the distance computation on Facebook. The network that was analyzed represents a true meaning of Big Data (approximately 69 billion friendship connections and 721 million users). More interestingly, they observed that the distance between two individuals on the network is 4.74 on average. This in turn corresponds to 3.74 degrees of separation, (i.e., number of links in between), among users. This finding is indeed interesting because it does not only confirm the previous findings but also demonstrates a new emerging phenomenon of a shrinking-world due to the worldwide use of social media.

We are interested in a special type of network based on keywords and hashtags gathered from tweets. We are greatly fascinated by the shrinking-world phenomenon exhibited on the Facebook network and presented by Backstrom et al. We are set out to investigate such a phenomenon in the networks of contents that are (1) gathered from a social media (i.e., Twitter) similar to Facebook and (2) support the social tagging mechanism enabling the use of hashtags. In our work, we use the

number of vertices of the paths found as a lower bound, and *eccentricity*, a vertex centrality metric, as an upper bound on the distance.

Our experiments and analyses have features similar to those by Dodds et al. and Backstrom et al. First, our K-H networks are constructed from data collected from Twitter, a platform similar to Facebook. Therefore, our networks are of a social-media origin even though it is not directly about users themselves but about what users say on this social media platform. Second, our experiments use keywords selected from four different domains (i.e., smoking substances, sentiments, drugs, and car models) which may or may not be directly related. This setup is similar to Dodd et al.’s setup of using different populations from 13 countries. Third, using 2 million tweets, we constructed two network representations: a keyword-keyword network and keyword-hashtag network. The generated networks produced 87,133 vertices (including the original keywords used to search Twitter) which are connected by 115,409 association links. This makes our networks similar in scale to the ones used by Dodds et al. Fourth, we allowed connections among keywords and hashtags whenever found, which makes our networks not of a bipartite type. This also ensures common grounds of similarities when comparing our analyses and results with the results presented by both Dodds et al. and Backstrom et al. Our research also overlaps with word association networks and their small-world properties. De Deyne et al. (Deyne & Storms, 2008) studied word association networks and observed that central nodes are highly frequent and are obtained early.

### 3.3 Definitions

**K-K Association Network:** Consider a set  $K$  of keywords extracted from tweets. A K-K network is a directed graph  $(V, E)$  where each node  $v \in V$  represents a keyword in  $K$  (i.e.,  $V = K$ ) and each edge  $e \equiv (k_i, k_j) \in E$  represents an association between the keywords  $k_i$  and  $k_j$  ( $i \neq j$ ). **K-H Association Network:** Consider a set  $K$  of keywords and a set  $H$  of hashtags extracted from tweets. A K-H network is a directed graph  $(V, E)$  where each node  $v \in V$  represents either a keyword in  $K$  or a hashtag in  $H$  (i.e.,  $V = K \cup H$  where  $K \cap H = \emptyset$ ) and each edge  $e \equiv (k, h) \in E$  represents an

association rule between a keyword  $k \in K$  and a hashtag  $h \in H$ . **Shortest Inter-keyword Distance:** In a K-H network, the shortest distance between two distinct keyword nodes is defined only between those that are not connected in the K-K network but are connected in the K-H network. So, it is measured as

$$\min_{k_i, k_j \in K' \wedge k_i \neq k_j} \{distance(k_i, k_j)\}$$

where  $K'$  is a set of the keywords that are connected in the K-H network while not in the corresponding K-K network, and  $distance(k_i, k_j)$  returns the path length (counted as the number of edges) between the two keyword nodes  $k_i$  and  $k_j$ . **Eccentricity of a Keyword:** Let  $k$  be a keyword node in a K-H network. The eccentricity  $ecc(k)$  of the keyword  $k$  is the maximum shortest distance from  $k$  to every other node, whether keyword or hashtag. That is,

$$ecc(k) = \max_{k \in K \wedge w \in K \cup H} \{shortest-distance(k, w)\}$$

where  $shortest-distance(k, w)$  returns the shortest path length from  $k$  to  $w$ .

### 3.4 Data Collection, Modeling and Network Construction

The data used in our study are from public tweets and are captured in real time. In addition to words, each tweet may contain such features as user handle, URL, and one or more hashtags. We excluded tweets that post pictures because they are out of the scope of this study. The features we are concerned with are words and hashtags. In order to collect tweet data, a user must specify search keywords to filter in the relevant tweets. Our datasets cover four different domains: (1) *cars* (manufacturer names of passenger vehicles, listed in the Kelly Blue Book (KBB, 1996)), (2) *drugs* (brand names of common over-the-counter drugs, obtained from the “drugs.com” (Com, 2001) site), (3) *sentiments* (common sentiments among people, obtained from the Facebook (Facebook, 2004) list of feelings embedded in the Update Status tab), and (4) *smoking* (terms referring to smoking substances common in the general public). As for the hashtags, they cannot be determined in advance.

They simply emerge in the tweets retrieved as a result of searching by the preselected keywords. Therefore, tweet words are related to the the embedded hashtags found in the same tweet. Hashtags can be of a free-form text that is comprised of one or more words preceded by the # sign. They do not follow a particular structure, while some people have used case (i.e., upper or lower) and special punctuation (e.g., underscore) to make them readable. Users can also choose to place multiple hashtags in the same tweet or simply send a hashtag as the entire tweet.

We collected five different datasets: one small dataset of about 50 to 500 tweets in each domain and one large dataset of more than two million tweets from all four domains together. Overlaps between domains are allowed. We used Twitter4J(Yamamoto, 2007), a Java Wrapper API, which connects to the official Twitter Streaming API to gather all five datasets. The rationale for using the four small datasets is to facilitate visualizing the datasets so that we can make observations and gain insights. Figure 3.1 shows the word-clouds associated with each small dataset.

Tweets are colloquial and informal and, therefore, noisy in nature. Any meaningful analysis must begin with eliminating the noise and constructing a data model. This process involves the removal of contents that are not in English. We detect identical words that are affixed with signs or special characters (e.g., “happy?”, “happy,”, “happy:”) and consolidate them. We also combine identical words that are in different cases (e.g., “acura”, “Acura”, “ACURA”). Once data are cleaned, each tweet is dissected to construct the following data models: (1) keyword-keyword occurrences, (2) keyword-hashtag occurrences, and (3) hashtag-hashtag occurrences. Each tweet is dissected to contribute its features to each of these three models whenever applicable. For instance, a tweet “I am so excited about the Big Data conference in Alaska #Big Data #2014 #scientists\_socialize” generates the following transactions: (1) *keyword-keyword*–(excited, ?); (2) *keyword-hashtag*–(excited, #ws14), (excited, #Big Data), (excited, #scientists\_socialize); (3) *hashtag-hashtag*–(#Big Data, #2014), (#Big Data, #scientists\_socialize), (#2014, #scientists\_socialize).

We construct the data models from each type of the transactions generated. Each corresponding data model is converted to the attribute-relation file format (ARFF) to be analyzed using the



Figure 3.1: Four sets of keywords used for small datasets.

Weka association analysis module (Frank et al., 2010; Hall et al., 2009). Using the Apriori algorithm (Agrawal & Srikant, 1994; Wu et al., 2007; Wu & Kumar, 2009), Weka produces rules from each model. When rules are generated from each corresponding model, they are saved in comma-separated values and used to construct the two types of networks, K-K and K-H. Depending on the minimum support and confidence thresholds specified in the experiments, Apriori may generate rules such as (#ws14, #btwon) and (#btwon, #ws14) which are identical pairs in opposite directions. In our work we disregard the direction of association as irrelevant and, therefore, prune out one of them. (Note, as a result, the association networks are undirected.)

## 3.5 Experiments

Experiments have been performed in two stages – first, in a small scale with keywords from each of the four domains. Second, a Big Data scale of millions of tweets that support the construction of networks for the given keywords. This dataset is a superset of all the four domains individually. The intuition is to find some supporting evidence of the shrinking-world phenomenon in the smaller dataset. Upon the emergence of the supporting instances, we test against a large-sale dataset to learn whether there will be improvement.

### 3.5.1 Small dataset experiments

In the K-H network of each domain, we investigate the paths connecting any pair of keywords that are not connected in the K-K network as shown in Figure 3.2. (Note that all keywords connected in the K-K network are naturally connected in the K-H network as well as demonstrated in Figure 3.3.) As expressed earlier, our particular interest lies in measuring the number of vertices of the two keywords if they are found connected in the K-H Network as a lower bound while using eccentricity as an upper bound. Some instances of such keyword pairs are shown in Figure 3.4. The keywords pairs are not connected in the K-K networks, seen in Figure 3.2, but are connected in the K-H networks, and most of them are connected via only one hashtag. The longest among all shortest paths found is between the keywords ‘loved’ and ‘excited’ in the sentiments domain, and its path length is seven vertices. This also means that particular instance is not a supporting evidence.

Now, further illustrations are provided for each domain and also shown in Figure 3.4.

**Cars:** In the K-K network, there are four pairs of keywords directed connected: (Honda, Acura), (Nissan, Mitsubishi), (Toyota, BMW), (BMW, Hyundai). In the K-H network, there are additional pairs of keywords connected, and most of these connections are via one hashtag. For instance, (Peugeot, #cars, Jaguar), (Jaguar, #memphis, Volvo), (Toyota, #tennessee, Volvo),



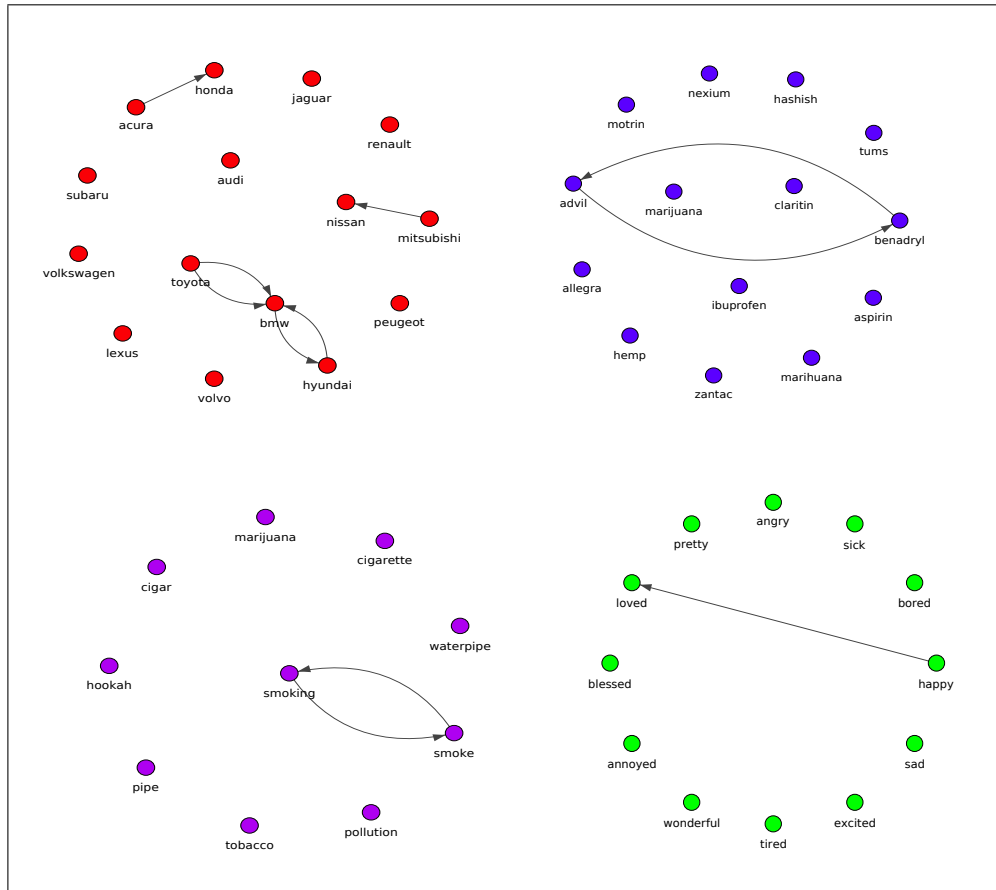


Figure 3.2: K-K networks of the four domains.

(Hyundai, #download, Lexus), (Lexus, #tennessee, Volvo), (Acura, #usedcars, Audi), and (Audi, #tennessee, Volvo). The maximum shortest path length between two newly connected keywords is 4 between peugeot and volvo.

**Drugs:** The K-K network shows only one pair of keywords directed connected, (Benadryl and Advil). The K-H network has additional keyword pairs connected, many other pairs directly via one hashtag. For instance, (Advil, #peopleschoice, Tums), (Benadryl, #peopleschoice, Tums), (Hemp, #mjnews, Marijuana) and (Hemp, #cannabis, Marijuana). The maximum shortest path length found is three, as demonstrated.

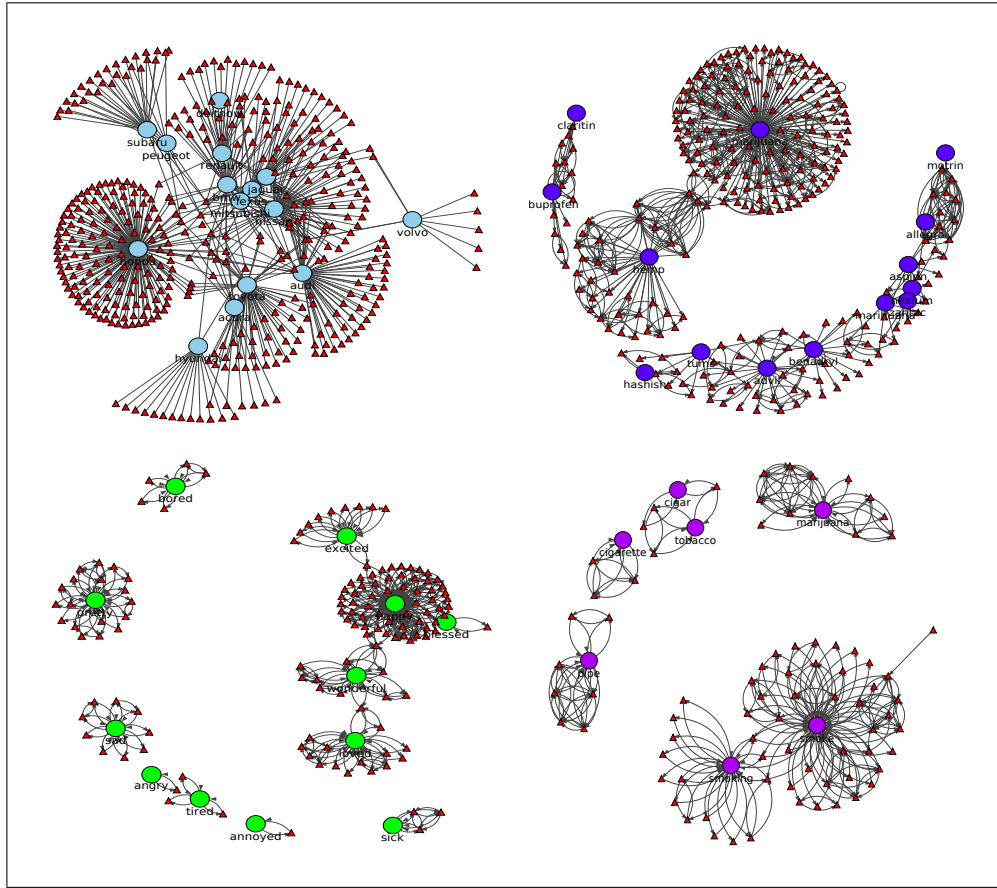


Figure 3.3: K-H networks of the four domains.

**Sentiments:** The K-K network has only one pair of keywords, (happy, blessed), connected. The K-H network shows additional keyword pairs connected (e.g., excited, happy). The maximum shortest path length is six in the path (loved, #xfactorfinal, wonderful, #gameinsight, happy, #bethanymotagiveaway, excited)).

**Smoking:** The K-K network shows a direct connection only between smoke and smoking. The K-H network shows additional keyword pairs connected via one or more hashtag. The maximum shortest path length is 3 in the path (cigar, #cringeworthy, tobacco).

In summary, we observed that some supporting evidence about keyword pairs in the K-H network presented in paths of three vertices: (cigar, #cringeworthy, tobacco), (marijuana, #cannabis, hemp),

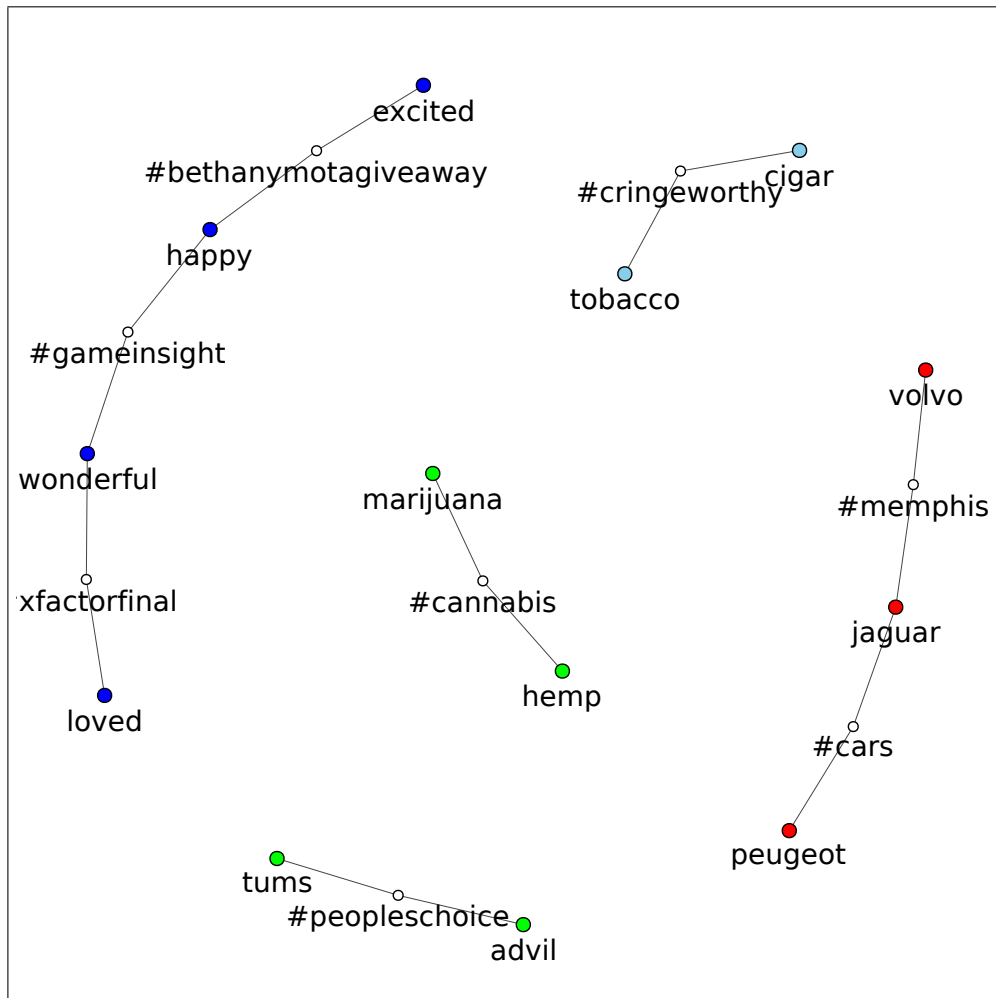


Figure 3.4: Some of the keywords connected in the K-H networks while not in the K-K networks.

(tums, #peoplechoice, advil). The eccentricity of each keyword vertex in all of the four networks fluctuated between (three and six). Based on these observations, the following questions have been raised for further investigation in a larger scale dataset.

- Will the number of vertices that make up the path of keyword pairs shrinks upon using a large-scale dataset?
- Will the eccentricity value improve as an upper bound also drop to become less than six?

### 3.5.2 Large dataset experiments

We further investigate the questions that we ended the small data section experiment. More than two million tweets will be used to answer the questions raised from the small dataset experiments, and the answers are affirmative. Figures 3.5 and 3.6 show portions of the K-K and K-H networks obtained from the large dataset. The K-H network is very dense, evidently showing a high level of interconnection among the domains made possible by the incorporation of hashtags.

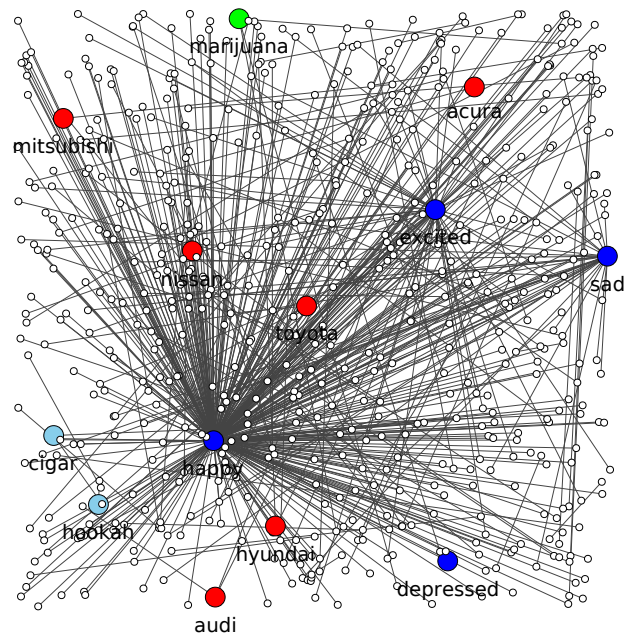


Figure 3.5: The four-domain K-H network.

### Analysis

Using means of a Random Walk (Noh & Rieger, 2004) algorithm, on the K-H undirected network, we randomly selected 40,798 paths of length four. We implemented a simple version of a walk where the algorithm is given an input vertex as a start point. The algorithm selects the neighbors

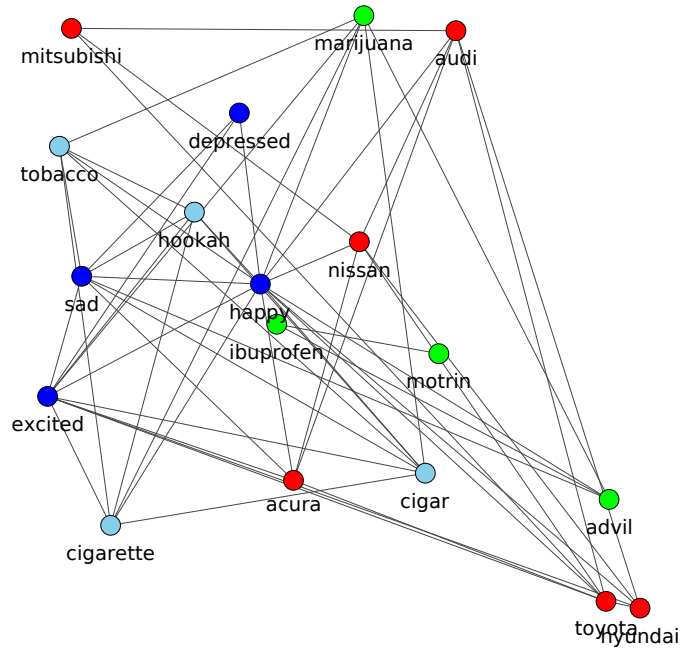


Figure 3.6: The four-domain K-K network.

of the vertex and continue to find neighbors until a walk of length four is achieved. The algorithm then moves on to another vertex until a reasonable size sample is gathered (i.e., 40,798 walks). We have obtained the following patterns along with their perspective frequencies (1) HHHH:4606, (2) HKHK:25346, (3) HKHH:7212, (4) HHKH:2057, (5) HHHK:1446, (6) KHHH:3, (7) KHKH:14.

Table 3.1 summarizes the patterns, frequencies and their percentages of all random walks generated. The table suggests that most frequent and dominant pattern is the HKHK with a 62% frequency. This also demonstrates that each individual keyword **K** is connected with another using at most one hashtag **H** (i.e., **K-H-K**). Additionally, this pattern suggests that hashtags tend to bond with keywords more frequently and in this alternation fashion. Another important observation from the patterns generated is the HHHH which scored 11% frequency. Though it is a rare chance to find four hashtags in a row, yet we still find a significant percentage of such a pattern. This suggests that hashtags can also bond with their own counter part but with a much less frequency. An overall observation about these patterns is that hashtags tend to stick to keywords more than to other hash-

tags. This is the common theme pattern among all patterns with the exception of the first pattern **(H-H-H-H)**. By measuring the eccentricity of keyword vertices in the K-H network, we have found that the eccentricity values for all keywords are exactly four. By measuring the eccentricity of the K-K network we found a trend of an ever increasing eccentricity value. This is due to the fact that some keywords are not connected to each other in the K-K network and the distance among each of pair of these disconnected vertices is infinity. Figure 3.7 shows the eccentricity percentile for both K-K and K-H networks.

Table 3.1: Various Keyword-Hashtag Patterns and Their Corresponding Frequencies

P1	P2	P3	P4	P5	P6	P7
HHHH	HKHK	HKHH	HHKH	HHHK	KHHH	KHKH
4606	25346	7212	2057	1446	3	14
11%	62%	17%	5%	3%	≈ 0%	≈ 0%

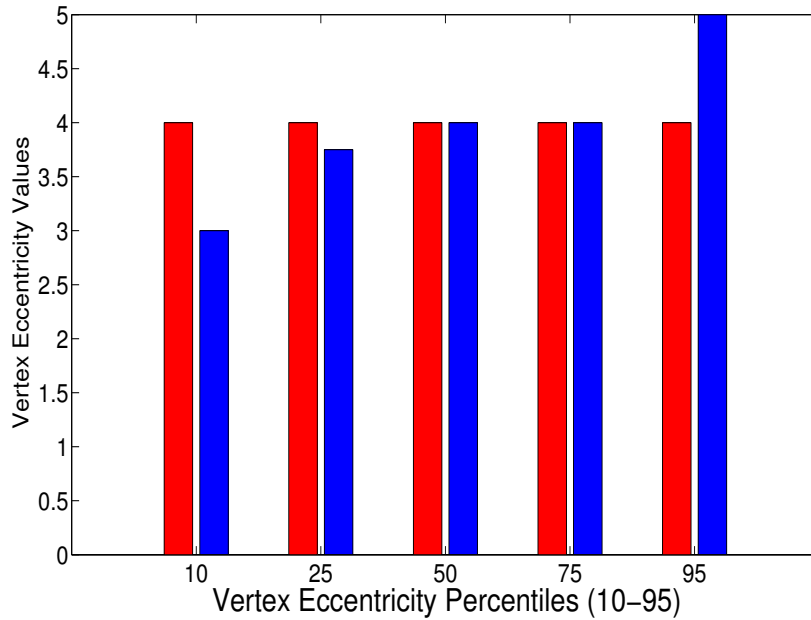


Figure 3.7: Eccentricity Percentiles

## Comparison with a random network

In order to have a more tangible observation of hashtag tendency to stick to keywords to the effect of shortening distances between them, we have measured the probability of edge insertion needed to achieve the same eccentricity value in a randomly generated network. For this purpose, we have selected the Erdős and Rényi's random network model (Erdős & Rényi, 1959). Figure 3.8 shows the results. It shows that in order to achieve the same eccentricity value (i.e., four) as in the four-domain K-H network, the random network needs the edge insertion probability of only 0.0008, whereas in order to achieve the same as in the four-domain K-K network, it needs the probability of 0.15, which is 187 times higher than that in the case of the K-H network.

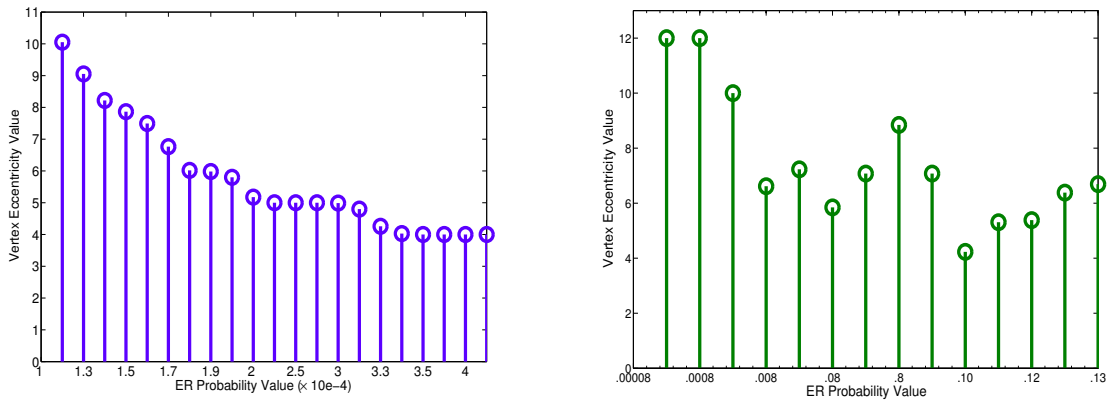


Figure 3.8: Vertex eccentricity measured from using Erdős and Rényi's random network model.

## 3.6 Discussion

Our study above further investigated an interesting phenomenon demonstrated in Big Data environments such as social networks. More specifically, we have presented further exploration of the (four degree of separation) findings of Backstrom et al. in 2012. The analysis was performed on networks of words and hashtags collected from tweets. The purpose of this study is to measure the impact

of hashtags on such networks and whether they contribute to the shrinking world observations of Backstrom et al. We constructed two different networks: The first network is from associations between domain keywords which we call the K-K network. The other is constructed from associations found among keywords and hashtags which we call the K-H networks. In this section, we provide the findings of the hypothesis and learned insights where knowledge discovery can be achieved.

The experiments of small individual datasets showed that: for some pairs of disconnected vertices in the K-K network, their paths are three to seven. We also measured the vertex eccentricity of each keyword in the network, which also fluctuates between three and six. In real world, however, such individual sets will not be found in isolation. To show the potential of the K-H networks, we blended the keywords of the four domains into one set. This is to enable the interconnections between keywords from different data sets with the possibilities of producing patterns such as (acura, excited), (smoking, sad), and (advil, marijuana). The large-scale experiments showed more interesting and supporting results to the shrinking world phenomenon observed by Backstrom et al. Finding that the number of vertices between each pair of connected keyword in the K-H networks is exactly three was indeed surprising but was also understandable when examining the random walk patterns. Having the eccentricity of each keyword is exactly four was a simple derivation from the fact that each pair of keywords is connected by exactly one hashtag. However, the eccentricity is an upper bound and there is always another hashtag vertex beyond the third vertex as a maximum shortest path in the network. Table 3.1 column P2 shows this exact case which also is the most frequent and dominant among all patterns.

We modeled the real world networks (i.e., K-K network and the K-H Network) using Erdős and Rényi's random model in order to estimate the probabilities of any pair of keywords to bond with each other with and without hashtags. Erdős and Rényi is the simplest to use since it generates an unbiased network contraction which makes it a good tool to use for comparison. The randomized networks that simulated the necessary probability values to produce eccentricity value of four for both the K-K network and K-H network. While the K-K network converged with a probability value



$p = 0.15$ , K-H network converged at  $p = 0.0008$ . This is clearly in compliance with our original observations produced by the random walk. This means that a very little probability is required to establish connections between two keywords when hashtags are incorporated. In the absence of hashtags, however, a much higher probability is needed to achieve the same results.

### **3.7 Conclusions**

In this study, we have investigated the contributions of Big Data environments to the shrinking world phenomenon. We have presented a new network model that we believe will facilitate knowledge discovery on social media. More specifically, we investigated the degrees of separation that was recently studied on a true meaning of Big Data and using a Big Data environment. Our network model proves supportive of the finding made by Backstrom and also proven a more shrinking world of contents gathered from Big Data environments. The number of vertices of the paths found were exactly three as a lower bound while eccentricity was exactly four for each keyword in the K-H network as an upper bound. This translates to a smaller degree of separation of only two links between each pair of keywords as opposed to 3.7 in the Facebook network of people measured by Backstrom et al. Our future direction will further investigate this phenomenon from Inter-Big Data environments where we gather tweets and Facebook posts to generate and study more complex K-H networks. One of the most promising aspects of presenting the K-H networks is to enable the knowledge discovery using means of Big Data environments. One of the interesting paths found was between (marijuana, ibuprofen), which was connected via the (#Alzheimer) hashtag. Upon reverse engineering the path and examining the original tweets, we found that a new study stated that Ibuprofen can be used to block memory loss caused by intaking medical marijuana. This study made its way to Twitter via some news organizations (e.g., LA Times). We believe that K-H networks can further expose other unknown relationships such as this one.

### **3.8 Acknowledgments**

The authors would like to thank Dr. Tamás Nepusz and The Python IGraph (Csardi & Nepusz, 2006) team for their tips and discussions on graph visualization.

## Chapter 4

# Biomedical Web Science: Mining Patterns in Concept-Hashtag Networks

Can Twitter hashtags yield clinicians a powerful tool to extrapolate patterns that may lead to development of new medical therapy and/or drugs? In our chapter , we present a systematic network mining method to answer this question. We present *HashnetMiner*, a new pattern detection algorithm that operates on networks of medical concepts and hashtags. Concepts are selected from widely accessible databases (e.g., Medical Subject Heading [MeSH] descriptors, and Drugs.com), and hashtags are harvested from associations with concepts that appear in tweets. The algorithm discerns promising discoveries that will be further explained in this chapter . The findings are linked to PubMed articles of the same subject. Our method offers a new way of bringing anecdotal information to Biomedical Digital Libraries like PubMed. To the best of our knowledge, this is the first Biomedical Web Science that addresses such question by means of data mining and knowledge discovery in hashtag-based networks.

## 4.1 Introduction

Social media sites generate seemingly unlimited types of knowledge and open a new research domain for social scientist, physicists, computer scientists, mathematicians, biologists, and medical scientists (McCarthy et al., 2013). Twitter, one of the most well known social media outlets, has attracted the attention of scientists more than any other social media. The various characteristics of Twitter feeds, also known as tweets, make these feeds appealing for research. A tweet may contain words, links to external websites, an image, and/or a special feature called a “hashtag”. A hashtag can be comprised of one or more words, some of which are separated by underscores, and others by upper-lower case convention (e.g., #social\_care vs #SocialCare). Hashtags serve to group tweets of a specific topic and are an invitation to join conversations. These free-form hashtags are extremely useful when analyzed in light of domain specific vocabulary.

Among the various usages of hashtags are ones that signal an event (e.g., #ItalyVsBrazil), to invite people to an event (e.g., #networking), or to recommend a product (e.g., #eCig). The emergence of these hashtags on Twitter marks the beginning of an era full of unexpected discoveries. The past few years have shown an ever-evolving usage of hashtags that encompasses a wide range of knowledge. Each individual hashtag functions almost like a neuron in the human brain, sending a specific signal and performing a distinct task. Among a large spectrum of signals, one may point to a location (e.g., #US), an outbreak (e.g., #HIN1), disease (e.g., #cancer), synonym (e.g., #weed, e.g., #pot to refer to Marijuana). Clearly, there are infinite types of hashtags.

Hashtags have become a core feature in almost every scientific article that makes its way to Twitter. Top scientific journals (e.g., Science Magazine, Nature Magazine) have finally found this much needed outlet to broadcast the briefings of newly published articles. The following are significant scientific discoveries that were tweeted, some on the Science Magazine’s Twitter account, and another was posted on the Nature Magazine’s Twitter account (“@sciencemagazine, and @NatureMagazine”):

- “Upon internalization, #Salmonella choose between replication or a form of quiescence <http://bit.ly/1hBQV8D> #microbio”
- “Painkillers May Curb Memory Loss From Medical Marijuana — Science/AAAS — News <http://bit.ly/1jf6Ryn> #alzehimer #alz”
- “This week’s #OutlookCancerImmuno covers the advances in utilizing the body’s immune system to fight cancer <http://bit.ly/18U9jUu>”

From the previous examples of tweets, it is evident that hashtags (i.e., #alzheimer, #OutlookCancerImmuno, and #microbio) play a crucial role in disseminating and connecting nuggets of information posted on Twitter. Gaining good understanding of how hashtags function is bound to reveal a wealth of knowledge. Regardless of how, studying these functions to understand their semantics is a very difficult and labor intensive task. It is imperative to automate these human intensive tasks to eliminate the human labor involved.

Developing a computational method to identify significant hashtags requires a great deal of creativity. Such a method must be comparable to establishing sound and measurable theories. In the recent years, the study of hashtags have proven to be appealing to scientists. Lehmann et al. focused their study on hashtag peaks which lead to discover spikes of collective attention on Twitter. When popular hashtags are identified, they are linked back to their original tweets for further analysis (Lehmann et al., 2012). Romero et al. studied the hashtag spread on a network defined by the interactions among Twitter users. The study found significant variations in the ways that widely-used hashtags spread in the network (Romero et al., 2011). Conover et al. studied hashtags as an important feature for clustering political polarization on Twitter (Conover et al., 2011). Wang et al. proposed a model that analyzes the traffic patterns of the hashtags gathered from streaming tweets, to generate adaptive subsequent collection queries (?, ?). Carter et al. designed a framework for tracking various aspects of worldwide events (e.g., earthquakes, political elections, etc) via the use of hashtags (#) (Carter et al., 2011). The framework extracts the list of hashtags associated with each

topic and accounts for multilingual contents. The their study used means of Information Retrieval to translate hashtags.

Studying hashtags in the light of other keywords in the same tweet is promising. The previous example (“Painkillers May Curb Memory Loss From Medical Marijuana — Science/AAAS — News <http://bit.ly/1jf6Ryn> #alzehimer #alz”) show that there is a connection between Marijuana and Ibuprofen via the #Alzheimer hashtag. The same tweet links to an article that presented evidence that Ibuprofen may block memory loss as a result of intaking medical Marijuana as a pain-killer. This very important linkage between (Ibuprofen and Marijuana) does not appear in previously published articles archived in PubMed. To complicate the matter even more, the article states a claim which contradicts a previously published article by Gorsky: (“Marijuana test: no ibuprofen interference”) published in the Science Magazine in 1989 (Gorsky, 1988), and archived in PubMed (PMID:3043663).

We believe that such discoveries and new evidence need to be incorporated into previously published articles via signifiant hashtags such as the (#Alzheimer and #Salmonella) in the above example. Systems that “tag” content with keywords are known as Academic Social Tagging Systems(e.g., Connotea and CiteULike), and they provide researchers with means of content organization using keywords (tags) (Good et al., 2009). The use of keyword metadata Medical Subject Headings (MeSH) descriptor has demonstrated promising in linking PubMed to other resources. Chen et al. performed a feasibility study to link molecular sequences and clinical trials using MeSH (Chen & Sarkar, 2010). Tasneem et al. (Tasneem et al., 2012), developed and validated a methodology for annotating studies employing Medical Subject Heading (MeSH) terms applied by an the National Library of Medicine (NLM) algorithm. None of these previously described systems studied newly emerging evidence using associations between MeSH and Hashtags. Hashtags are special devices and carry much knowledge within as we will demonstrate in the Preliminary section.

Our main goals in this article are to do the following:

[noitemsep,nolistsep]Show the *significant types of discoveries* that hashtags can transpire when studied in the light of vocabulary of interest. Perform a systematic network mining using HashnetMiner algorithm that detects path patterns and generates knowledge from a hashtag-based network Link significant patterns vetted by clinicians to related PubMed using MeSH equivalent terms.

## 4.2 Data

This study uses two different types of data resources:

[noitemsep,nolistsep]**Search Keywords** that are relevant to the domain of interest. In order to crawl Twitter using the APIs for any set of tweets a user must identify a given set of search keywords. Twitter APIs in turn retrieve any tweet that has one of more of the search keywords. The more relevant the keywords to the domain, the likely the tweets will support the hypothesis behind the search task. For the preliminary experiments we used terms from MeSH related to Tobacco and Marijuana smoking. For the network mining tasks, we used a combination of MeSH terms and terms available in Drug.com. We found that MeSH terms alone are too technical and does not retrieve much of Twitter traffic when used for search. Therefore, we used 60 drug brand names which were mapped to MeSH using their drug generic. A full list of terms is available at the smoking/marijuana MeSH branches: (1) “F01.145.466.349.750.488”, (2) “C25.775.635”, and (3) “B01.650.940.800.575.100.175.500”. Figure 4.1 shows a Word-Cloud for all the drug brand names used in this chapter . **Public Tweets** that are gathered using the preselected search keywords. Public tweets capture current events posted by individuals from all over the world. Each tweet is comprised of various components such as words, hashtags, user handles, URLs and other components which make up the entire tweet. The preliminary dataset used 25 tweets only related to tobacco smoking to show the various types, and listed in Table 4.1 of hashtags and their significance. The network mining tasks performed on a set of 2 million tweets to identify patterns that can lead to marijuana-drug





No.	Original Tweet Text
1	r u ready to quit? #stoptober starts in 5 days. visit us for help and advice on stopping smoking
2	#didyouknow if you stop smoking for 28 days, you're 5 times more likely to stay smoke free. join
	in with #stoptober
3	anyone giving up smoking this month? #youcandoit #stoptober
4	mhm is supporting stoptober this year, only smoking going on here will be from our natalies
	exhaust. #stoptober lets get healthy. :)
5	quest completed matches for smoking pipes, #android, #androidgames, #gameinsight
6	construction completed tobacco shop, #android, #androidgames, #gameinsight
7	i managed to finish the cigarette smoke assignment! try it for yourself! #gameinsight #android
	#androidgames
8	Smoking #hookah in #Istanbul! #ttot #travel #rtw #lp #turkey
9	Medium contract is now available at Tobacco Shop, #Android, #AndroidGames, #GameInsight
10	Now Tobacco Shop has reached level 2 , #Android, #AndroidGames, #GameInsight
11	I got the most smashed I ever been on this day... really smoking a cigar at the range... #TBT
12	#tbt #halloween #2010 ? cigarette girl... This has still been my favorite costume yet
13	#tbt 1 of my favorite non smoking nice young man pics lol my godson took it when he was like 3
14	Throwback to tobacco harvest chevytravis #tbt
15	#tbt My bachelor party. Cigar & Jameson while the dudes get ready. #atlanticcity #suitelevel
16	#tbt smoking hookah. .. hookah is stupid btw
17	#tbt hookah bar fun
18	#TBT my bro @LF605 used 2 roll dem limos up...dis nigga gotta start smoking again lmaooo
19	RT @CPME_EUROPA: Congratulations #ExSmokers please help deter children & young people
	from taking up smoking #tobacco #TPD
20	RARE Signed Kaywoodie Tobacco Pipe Orange White BAKELITE ? Clover & Rhinestone \$29.99
	#tobacco #cigarette
21	Hardest body attack sesh. Gotta quit smoking #Yolo #smoking #Iworkout
22	Check out our Dual Pro Electronic Cigarette starter kit! We also have a menthol flavored kit #ecig
	#smoking #menthol
23	Smoking Linked to \$278 Billion in Losses for U.S. Employers #tobacco
24	RT @CPME_EUROPA: Congratulations #ExSmokers please help deter children & young people
	from taking up smoking #tobacco #TPD
25	#Obama: I Quit Smoking Because I'm Scared of My Wife via @mashable

Table 4.1: Dataset2 – Real-World tweets

### 4.3.1 Proof of Concept

To assess the usefulness of hashtags, we used a toy dataset that is comprised of 25 tweets to test our initial intuition. To eliminate any bias in the initial experiments, we used Twitter Live Streams which returns a 1% sample of the entire tweets from all over the world. In order to retrieve any tweet from the Twitter Streaming API, a set of keywords must be entered as an input. There is no guarantee that the tweets resulting from the query will have hashtags. If hashtags exist, some may be identical to the keywords used for querying the Twitter feed, others may not. A tweet such as “I need to quit smoking cigarettes #tobacco #nicotine” does indeed have two instances of the input keywords Smoking, and Tobacco. Additionally, there are two different types of hashtags: “#tobacco” which is identical to the input keyword Tobacco and “#nicotine” which is emergent and is not among the search keywords used.

Following Agrawal et al. (1994), and others (Creighton et al. 2003, Becquet et al., 2005), we view the creation of an association as a *co-occurrence* between two keywords in the same tweet. The Apriori algorithm of Agrawal et al finds new associations between items that co-occurred in the same transaction. Similarly, we consider each tweet as a transaction of its own, such that, item *A* is a keyword that exists in one of the preselected MeSH Smoking branch and item *B* is a hashtag that is either identical to MeSH term or entirely emergent. To construct a database of transactions similar to one of Agrawal et al., we follow a specific procedure: upon reading a tweet, its content is dissected to extract keywords, identical hashtags and entirely new hashtags. Co-occurrence belongs to each type make up the transactions. There are three different types of records to be constructed: In **First Type**: we construct transactions of taxonomy keywords that co-occurred together. For instance, if a tweet comes as “Smoking is such a hard habit to give up, who invented cigarettes or hookah? #cig #hubblyBubbly #hookah”. This tweet results in the following transactions (smoking, cigarette), (smoking, hookah), and (cigarette, hookah). Each transaction occupies a record in a Comma Separated Value (CSV) database. The **Second Type**: is strictly for transactions formed by keywords co-occurring with identical hashtags. The hypothetical tweet above results in

Table 4.2: Transaction Types and Samples. Records with “?” Indicate Missing Values

Term to Term Trans	Term to Identical hashtag Trans	Term to Emergent Hashtag Trans
(smoking,tobacco)	(tobacco, #smoking)	(tobacco, #nicotine)
(smoking, ?)	(tobacco, #cigarette)	(tobacco, #nicorette)
(smoking, cigar)	(hookah, #smoking)	(smoking, #420)
(cigar, tobacco)	(smoking, #smoking)	(tobacco, #ecig)
(smoking,tobacco)	(cigar, ?)	(smoking, #mmot)
(smoking, tobacco)	(tobacco, #cigarette)	(tobacco, #research)
(hookah, smoking)	(cigar, #tobacco)	(hookah, #istanbul)

the following record: (hookah, #hookah) **The Third Type:** is for transactions that are formed from the co-occurrence of keywords and emergent hashtags: (Smoking, #hubblyBubbly), and (Smoking, #cig). The database is then converted to Weka’s Attribute-Relation File Format (ARFF) (Hall et al., 2009) for association analysis using the Apriori algorithm. Table 4.2 shows sample transactions of each type that was generated from the tweet dataset.

### 4.3.2 Emergent Hashtags Rule Discovery

Tweets not only contain identical hashtags but also contain ones of different semantics and syntax. The rules harvested from this type of experiment can introduce new findings in virtually any domain. Emergent hashtags are surprising because we only know one side of the rule which is the MeSH term. As for the other side, it is unknown as it emerges from the themes propagating on Twitter. This basic idea is extremely powerful because it associates and connects various real-time context types with the vocabulary under examination. Figures 4.2 show four drugs namely (tums, ibuprofen, advil, protonix) with two transactions on the left side, while the right side shows that the same transactions

when (#relief) hashtag, showing in green, is introduced. A single hashtag has made all network nodes of four drugs connected. The type of knowledge that can be revealed using hashtags are abundant. This simple idea that can unleash a never-ending flow of knowledge. Table 4.2 showing the transactions generated from associations of MeSH terms and emergent hashtags.

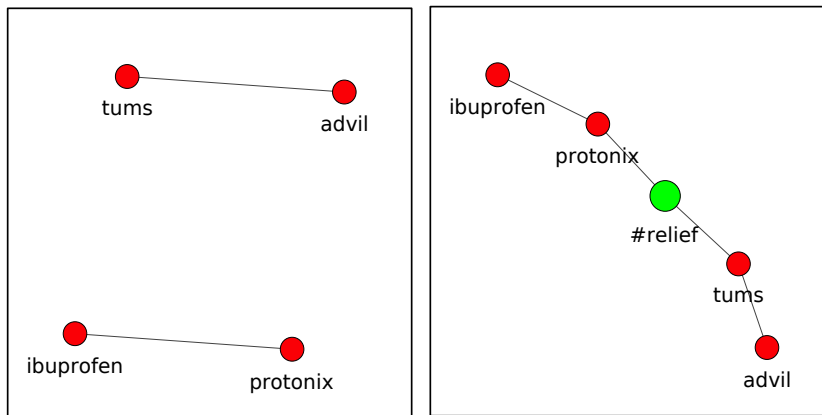


Figure 4.2: Impact of Hashtags When Introduced to Disconnected Networks

### 4.3.3 Types of Hashtags Found

The following are the various types of hashtags that were generated from the small dataset of 25 tweets. For the sake of enumerating some knowledge types hashtags can produce, we set the minimum support and minimum confidence to 2% and 100% respectively. This low support guarantees the inclusion of many hashtags regardless of the how frequent they are in the given rules. The network mining described in Section 4.4 explains how we prune insignificant rules. Figure 4.3 shows the association rules generated from a 25 tweet dataset by analyzing the the keywords when hang-tags are incorporated. In the following section, we present hashtag types, meanings, and significance in the Twittersphere:

[noitemsep,nolistsep]**Activity**– Five different discovered hashtags suggested an activity type: The first three (#TTOT which is an acronym for Travel Talk on Twitter, #RTW an acronym for Round the World, and #travel) which were linked to the MeSH Smoking concept and travel.

Figure 4.3: Association Rules mined by running Apriori against a small dataset of 25 tweets

```
Apriori
=====

Minimum support: 0.02 (1 instances)
Minimum metric <confidence>: 1
Number of cycles performed: 20

Generated sets of large itemsets:

Size of set of large itemsets L(1): 31

Size of set of large itemsets L(2): 38

Best rules found:

 1. hashtags=#stoptober 4 ==> kwds=smoking 4 <conf:(1)>
 2. hashtags=#exsmoker 2 ==> kwds=smoking 2 <conf:(1)>
 3. kwds=hookah 2 ==> hashtags=#tbt 2 <conf:(1)>
 4. hashtags=#didyouknow 1 ==> kwds=smoking 1
 5. hashtags=#youcandoit 1 ==> kwds=smoking 1
 6. hashtags=#hookah 1 ==> kwds=smoking 1 <conf:(1)>
 7. hashtags=#istanbul 1 ==> kwds=smoking 1 <conf:(1)>
 8. hashtags=#ttot 1 ==> kwds=smoking 1 <conf:(1)>
 9. hashtags=#travel 1 ==> kwds=smoking 1 <conf:(1)>
10. hashtags=#rtw 1 ==> kwds=smoking 1 <conf:(1)>
11. hashtags=#lp 1 ==> kwds=smoking 1 <conf:(1)>
12. hashtags=#turkey 1 ==> kwds=smoking 1 <conf:(1)>
13. hashtags=#TPD 1 ==> kwds=smoking 1 <conf:(1)>
14. hashtags=#Yolo 1 ==> kwds=smoking 1 <conf:(1)>
15. hashtags=#lworkout 1 ==> kwds=smoking 1 <conf:(1)>
16. hashtags=#obama 1 ==> kwds=smoking 1 <conf:(1)>
17. hashtags=#halloween 1 ==> kwds=cigarette 1 <conf:(1)>
18. hashtags=#2010 1 ==> kwds=cigarette 1 <conf:(1)>
19. hashtags=#ecig 1 ==> kwds=cigarette 1 <conf:(1)>
20. hashtags=#menthol 1 ==> kwds=cigarette 1 <conf:(1)>
21. hashtags=#cigarette 1 ==> kwds=tobacco 1 <conf:(1)>
22. kwds=cigar 1 ==> hashtags=#tbt 1 <conf:(1)>
```

The fourth, showed that the#suitelevel associated with hashtags, refers to a social activity that takes place in the Suite Level of a stadium during a ball game. This indicates a causality relationship that leads to heavy smoking activities. The fifth, #IWorkOut, was also found connected to the smoking keyword. Referring back to the original tweets, we can state that smokers who want to quit, choose to work out in order to recharge their themselves with good

energy and overcome the cravings. **Date/Time**– We found two different hashtags that referred to year #2010 which were linked to cigarette. Particularly, this hashtag refers to a personal experience that took place in 2010 and involved a cigarette. The #TBT hashtag refers to a day of the week which stands for (Throwback Thursday), but also refers to a call for an event. The hashtag was associated with Hookah which indicate a social event that will involve hookah.

**Geolocations**– Three different geolocations were captured by hashtags: #Atlantacity, refers to Atlanta, Georgia which was linked to cigar, and the two others #Turkey and #Istanbul which refer to the country of Turkey and its capital Istanbul. The two hashtags referred to a smoking event and revealed its geolocation. **Holiday**– One hashtag (#Halloween) referred to a holiday and linked it to cigarette terms. This reveals that perhaps smoking rates rise around seasons or holidays such as Halloween, Thanksgiving or other. **Acronyms**– Two acronyms have been exposed in relation to smoking (1) #TPD is referring to the Tobacco Product Directive (2); #YOLO an acronym which stands for “You Only Live Once”. The tweets revealed that people use motivational slogans such as this one to give up smoking or social habits that might affect their quality of lives. **Sentiment**– Just as stated above, another slogan was also found linked to smoking as a positive sentiment #youCanDoIt. Sentiment analysis on Social Media and learning all positive and negative sentiments when studying a specific domain would be of a great use. **Substance**– The hashtag #Menthol was associated with cigarette, which reveals a very special type of community of smokers who are addicted to Menthol Cigarettes. **Online Service**– Two different hashtags pointed to two very different online services. One service offers products for quitting tobacco to smokers called Stoptober. The (#LP) is an acronym for “Lonely Planet“ which is also an online service specialized in making travel plans, provides information about lodging, resorts, cultures and food places. This information is very useful if it is compiled in a directory for consumers who may be interested in learning about options or businesses to learn about their competitors.

## 4.4 Method

The preliminary work presented above has proven to reveal various and unlimited types of hashtags using Association Rule Mining. The meaningfulness of each association is merely dependent upon the semantics of the hashtags. Clearly, this requires a great amount of human labor to filter out the important rules. Therefore, an automatic approach is desperately needed to overcome these labor-intensive efforts. In this section, we extend the ordinary association rules to association networks. We then present a systematic network mining method of identifying and ranking the rules according to strength using a newly designed algorithm called *HashnetMiner* for path mining. The steps and the algorithm description is listed and described as following:

[noitemsep,nolistsep]Association analysis and network construction A Network Mining using HashnetMiner a path mining and pattern detection algorithm Mapping and linking to PubMed

### 4.4.1 Association Network Construction

From the dataset of  $\approx 2$  million tweets we gathered using the Twitter streaming APIs, we generated three data models (1) *Keyword Co-occurrences*, from the words that appeared in the same tweet and have an exact match with the MeSH terms selected above. (2) *Keyword-Hashtag Co-occurrences* that were generated from the incidents of MeSH terms appearing in the tweet and the hashtags that also co-occurred within the same tweet. (3) *Hashtag Co-occurrences* from the incidents of all hashtags co-occurred in the same tweet in the same manner described previously in Table 4.2. Similar to the preliminary steps, we used the Apriori algorithm to generate association rules from each of the three data models we generated. This ensures that we have a network that can exhibit links between two keywords, a keyword and a hashtag, and two hashtags. This in turn can expose paths between two keywords that can not be shown without hashtags involved. This also means that the network generated from the associations is not bipartite. Upon the generation of the association rules, we perform further noise removal from the rules that involve hashtags. This is due to the fact

that hashtags are of free-form and could be entirely made of symbols that does not have particular semantics. Hashtags such as (“#?”, “#??”, “#\$\$%\$&”) are indeed valid hashtags, however, they do not carry any useful information within. Therefore, we prune out the entire rule that is comprised of such hashtags. Additionally, we pruned out rules that consists of hashtags of a length less than two characters (e.g., #9, #f, #1). Further validation was performed to remove all other invalid transactions that may have contained more than two data items due to parsing errors.

We store all other remaining rules in CSV files to generate two association networks: (1) a network from the drugs alone which have occurred in the tweet, and (2) the other network is comprised of all components and associations returned by Apriori. Using the iGraph Python Library (Csardi & Nepusz, 2006), a network can be generated from the associations rules exist in the CSV files. Since association rules are directed in their own nature, the network generated is also directed. Further simplification is performed to convert the network to an undirected network. Additionally, all redundant edges are removed. The iGraph Library ensures that identical keywords or hashtags have the same unique identifiers and are treated as single nodes. Spelling variations of the same keywords are treated as different vertices. For instance, the keyword (marijuana) and (marihuana) are treated as two different vertices and not consolidated. This is for various reasons: (a) to eliminate the labor-intensive tasks that are done by humans, (b) systemically prune out the rules that are not proving significant, and (c) enable linkage between terms that have similar semantics and explore their connectivity.

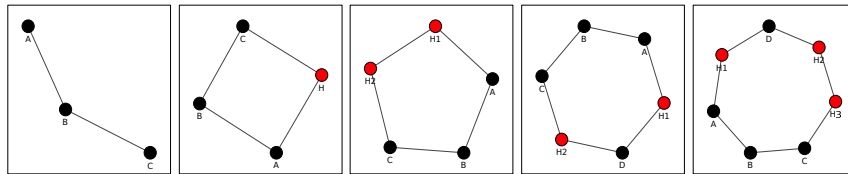
#### **4.4.2 Network Mining Heuristics**

Hashtags have proven to be powerful in exposing hidden paths between terms. Thus, it is essential to design heuristics that are based on this particular feature of hashtags. Allowing hashtags to establish such paths will indeed be a useful heuristic. Hidden paths can be exposed in various patterns. We select two particular patterns to minimize the search space: **(1) Pairs of any disconnected keywords** in the keyword association network, and search if there exists a path when hashtags are incorporated



(2) **Triadic Closure principle**, which is the property among three or more nodes A, B, and C, such that if a strong tie exists between A-B and A-C, there is a weak or strong tie between B-C (Watts, 2004). This particular property is the core mining heuristic (Rapoport, 1953; Boshmaf et al., 2011) that *HashnetMiner* uses to identify significant findings from the given tweets. By searching the network of terms for open triads, we generate possible candidate patterns one pattern at a time. Upon the identification of a pattern, the algorithm contrasts the other network (i.e., Concepts-Hashtags), to identify whether some the two patterns are matched. *HashnetMiner* is designed around the notion of “search and contrast” two graphs. This idea was computationally leveraged by Ting and Bailly in 2006 (Ting & Bailey, 2006) for mining and contrasting subgraphs to capture the structural differences between any two collections of graphs. Figure 4.4 shows an open triad formed only by terms found in tweets. The remaining subfigures of shows the possible ways how the original open triad is closed when hashtags are involved.

Figure 4.4: A, B, C Nodes form an Open Triad, and Possible Closures Using Hashtags or MeSH Terms Combinations



*HashnetMiner* searches and contrasts all open triads that have at most one hashtag in between. If K is satisfied, the process terminates and no further patterns are generated. In the case of not getting the required set of results, the process continues to searches for triads that are closed by at most two or three or fours intermediate vertices. The triads that were found closed are the findings and the final outcome of this process.

Our premise for finding the shortest paths of disconnected terms and closed triads is based on the principle that (“actors that are connected at short lengths or distances may have stronger connections”) (Hanneman & Riddle, 2005). Searching for the triads that are closed by at most one incrementally and up to four, reveals the hidden paths among the terms and may reveal interesting

findings. Further, the algorithm uses this premise as a ranking mechanism which in turn treats patterns with shorter length as stronger findings while longer paths counterparts are weaker findings. The outcome is a collection of patterns ranked according to their strengths in a descending order (i.e., the strongest first). The algorithm run analysis is:  $O(n^3 + m)$  according to Floyd-Warshall's algorithm (Hougary, 2010), where  $n$  is number of vertices in the graph and  $m$  the number of ranked patterns found. Algorithm 6 describes the steps of identifying significant rules by exploring the lengths of the closed triad of the entirely disconnected terms in the hashtag-based network.

---

**Algorithm 6** *HashnetMiner* is a pattern detection algorithm that contrasts two graphs for given patterns.

---

**Input:**  $K$ : desired number of patterns to be found

$G$ : a graph of keyword association

$G'$ : is a superset of  $G$  with hashtags

**Output:**  $K_p$  patterns discovered

**Description:**

```

1: Foreach vertex  $v_i \in G$ 
2:   Foreach vertex  $v_k \in G$ 
3:     If  $v_i \neq v_k$ 
4:       path  $p \leftarrow \text{compute\_shortest\_path}(v_i, v_k)$ 
5:       ;;path contrasting and mining step
6:        $\forall \text{length}(p) = \infty \vee \text{length}(p) > 3$ 
7:       path  $G'_p \leftarrow \text{compute\_shortest\_path}(v_i, v_k)$ 
8:       If  $G'_p \text{ length} \leq 3$ 
9:          $K_p \leftarrow K_p \cup G'_p$ 
10:      End If
11:    End If
12:  End For
13: End For
14: rank( $K_p$ )
15: Return  $K_p$ 

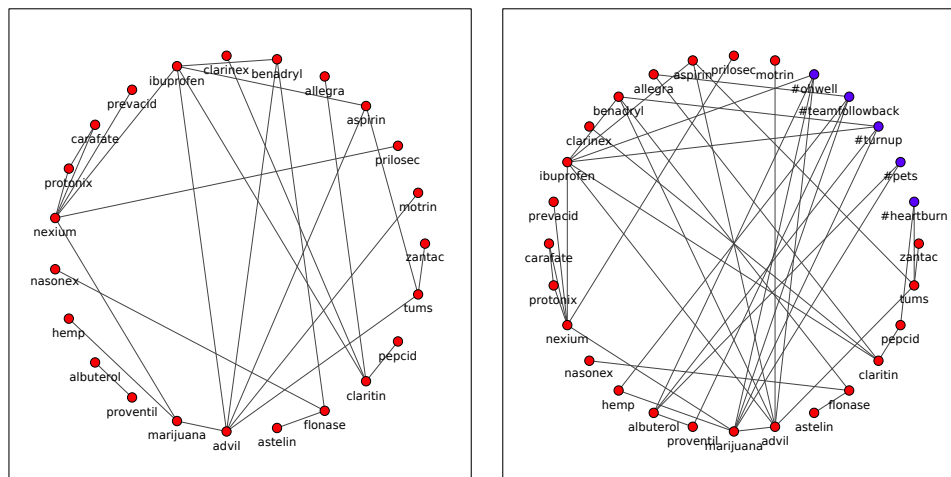
```

---

Figure 4.5 shows the newly discovered paths when hashtags are introduced. The network to the left shown in all drugs associations before pattern matching algorithm is introduced while the one to the right shows the discovered paths after. This network has 23 vertices and 26 edges. The network to the right shows that impact of the patterns discovered due to the introduction of hashtags showing

in blue. This network has 28 vertices and 42 edges. The hashtags introduced 5 new vertices that were not part of the original network. The algorithm discovered 22 closure patterns which introduced 22 new edges to the original network.

Figure 4.5: Impact of Hashtags When Introduced to Disconnected Networks. Top figure shows the drug associations when no hashtags are present. Bottom figure shows closed patterns when hashtags are incorporated



#### 4.4.3 Mapping to MeSH and Linking to PubMed

Inspired by the convention of hashtagging newly published articles posted on Twitter, we pose the following question: Can we hashtag Medline literature? Hashtagging a tweet is an equivalent mechanism for MeSHing PubMed articles. Both mechanisms are important yet they are entirely disconnected. This step is to translate the drug brand names and map them to their generic drug MeSH terms. Additionally, we link each generic drug names to its Registry Number RN/EC. RN is an identifier representing the substances mentioned in PubMed article. The RN field may contain any of the following: (1) the unique 10-digit Unique Ingredient Identifiers (UNII) assigned by the Food and Drug Administration (FDA) Substance Registration System (SRS). (2) the 5- to 9-digit number in hyphen separated format which is assigned by the Chemical Abstracts Service (CAS), and (3) for enzymes, EC number derived from Enzyme Nomenclature (NLM, 2005a). Such identifiers

Table 4.3: Mapping Drug to Generics and Linking to MH and EC/RN

Drug Name	Generic	No. MH Recs	EC	No. CoOccurr
albuterol	albuterol	8543	QF8SVZ843E	600
marijuana	marijuana	6555	–	–
benadryl	diphenhydramine	3754	8GTS82S83M	337
zantac	ranitidine	5907	884KT10YB7	757
prevacid	lansoprazole	1799	0K5C5T2QPG	0
tums	calcium carbonate	5008	H0G9379FGK	629
pepcid	famotidine	1455	5QZO15J2Z8	212
protonix	pantoprazole	0	–	–
prilosec	omeprazole	8353	KG60484QX9	478
clarinex	desloratadine	0	–	–
nasonex	mometasone	0	–	–
nexium	esomeprazole	650	N3PA6559FT	13
prednisone	prednisone	33982	VB0R961HZT	4623
allegra	fexofenadine	0	–	–

are not always included in the MeSH record for the substance. Hence, we have calculated the co-occurrences of each generic name appearing in in the MeSH field against the EC/RN field. Table 4.3 shows the mapping and linking of the drug brand names to MeSH and EC numbers.

## 4.5 Analysis

From the associations of experiments on the drug dataset, we found the following top 10 association rules (miniSup = 30% and miniConf = 65%) ranked according to their actual confidence.

- **(hemp, marijuana)**: it is expected since hemp is used a synonym for marijuana
- **(zantac, tums)**: two drugs are used for heartburn
- **(aspirin, ibuprofen)**: they belong to the same pharmacological class and have similar clinical indications
- **(prevacid, nexium)**: they belong to the same pharmacological class and have similar clinical indications

- (**clarinex, claritin**): they belong to the same pharmacological class and have similar clinical indications
- (**nasonex, flonase**): they belong to the same pharmacological class and have similar clinical indications
- (**albuterol, proventil**): they belong to the same pharmacological class and have similar clinical indications
- (**prilosec, nexium**): they belong to the same pharmacological class and have similar clinical indications
- (**pepcid, claritin**): they do not belong to the same pharmacological class. From the Twitter traffic we examined, it is possible that the same person were using both drugs. No signs of drug interaction was found from the tweets.
- (**allegra, claritin**): they are both used as anti-allergic drugs.

The *HashnetMiner* algorithm has produced 13 matches when searching for the two predefined patterns. With an exception of (tums, zantac), the algorithm has found novel and complex patterns that Apriori keyword association experiment could not find. Table 4.4 shows each match using its common drug names and the linked generic names.

Upon searching PubMed and Twitter for each pattern we found the following:

- (**Marijuana, #turnup, Albuterol**) : Albuterol is a sympathomimetic amine that can cross the blood brain barrier and induce central psychic effects. Tweeters are claiming that combining albuterol and marijuana would synergies for inducing euphoria. No scientific evidence is supporting this; however, an additive effect would be expected.
- (**Marijuana , #aids , Clarinex**) : The clinical use of cannabinoid receptor agonists in the treatment of AIDS symptoms may also exert beneficial adjunctive antiviral effects and regulate HIV infectivity(Costantino et al., 2012)

Table 4.4: Rules found when searching for disconnected vertices pattern and open triad patterns

<b>Common Term Pattern</b>	<b>Mapped Term Pattern</b>
(marijuana, #turnup, albuterol)	(marijuana, #turnup, albuterol)
marijuana, #pets, albuterol)	marijuana, #pets, albuterol)
(albuterol, #turnup, ibuprofen)	(albuterol, #turnup, ibuprofen)
(albuterol, #turnup, benadryl)	(albuterol, #turnup, diphenhydramine)
(zantac, #remember, prevacid)	(ranitidine, #remember, lansoprazole)
(tums, #heartburn, pepcid)	(calcium carbonate, #heartburn, famotidine)
(tums, #heartburn, protonix)	(calcium carbonate, #heartburn, pantoprazole)
(pepcid, #obamacare, hemp)	(famotidine, #obamacare, hemp)
(pepcid, #heartburn, protonix)	(famotidine, #heartburn, pantoprazole)
(pepcid, #heartburn, prilosec)	(famotidine, #heartburn, omeprazole)
(marijuana, #aids, clarinex)	(marijuana, #aids, desloratadine)
(nasonex, #prednisone, nexium)	(mometasone, #prednisone, esomeprazole)
(allegra, #food, prevacid)	(fexofenadine, #food, lansoprazole)

- (**Albuterol, #turnup, Ibuprofen**) : Common drug interaction is anxiety and cholelithiasis (gall bladder stone formation).
- (**Mometasone, #Prednisone, Esomeprazole**) : clinical pharmacists <sup>\*</sup>, <sup>†</sup> believe that these three drugs maybe used together in the treatment of Chronic Pbstructive Pulmonary Disease(COPD), bronchitis and other pulmonary conditions. This combination may already exits in the medical community. This may yield to further discussion.
- (**tums, #heartburn, pepcid**), (**tums, #heartburn, protonix**), (**pepcid, #heartburn, protonix**), (**pepcid, #heartburn, prilosec**) : all patterns belong to the same pharmacological class.

#### 4.5.1 Browsing Findings using a WebClient

When *HashnetMiner* detects the specified patterns, they are stored in a database in order to make it accessible for browsing. We developed a web interface where the end users can issue queries against this database and explore the linkage between any paris of drugs. The current application supports queries of the following type:

[noitemsep,nolistsep]**PMID query** – where the end user can issue a PubMed article ID(PMID) of a given article. The application responds by returning the relevant MeSH terms that were connected using one of more hashtags. The application supports hashtags linkage to Twitter. User can directly click on these hashtags and this action will triggers a search query against Twitter. The query find real-time traffic of this hashtags and what terms it connects to. **One or Two MeSH Terms Query** – In addition to searching for articles by their unique identifiers (PMIDs), the application supports searching for exact MeSH term or a pair of MeSH terms to check whether they are connected by a hashtags, and whether the connection is significant. When such a query is issued, the application responds by returning the all

---

<sup>\*</sup>Glen Myer RPH – Clinical Pharmacist.

<sup>†</sup>Tamer Fandy, Ph.D. – Assistant Professor Department of pharmaceutical sciences, Vermont Campus.

PMIDs that contain the MeSH terms of the query. Additionally, it finds all paths that explain how the terms are connected. User can also explore the hashtag real-time traffic by clicking on the hashtags in the same manner as above. Searching for (marijuana Ibuprofen) returns all PMIDs that contain marijuana and Ibuprofen. The query result shows that some of the hashtags connected the terms are (#Alzheimer). Figure 4.6 shows a screenshot when such a query is performed and the hashtag is displayed. **Hashtags Query** – The application also enables searching for hashtags. Whenever such a query is issued the application shows the MeSH terms that are connected by the hashtags. All PMIDs of the articles related to this hashtags are also returned. Each PMID are clickable which makes exploring the literature possible

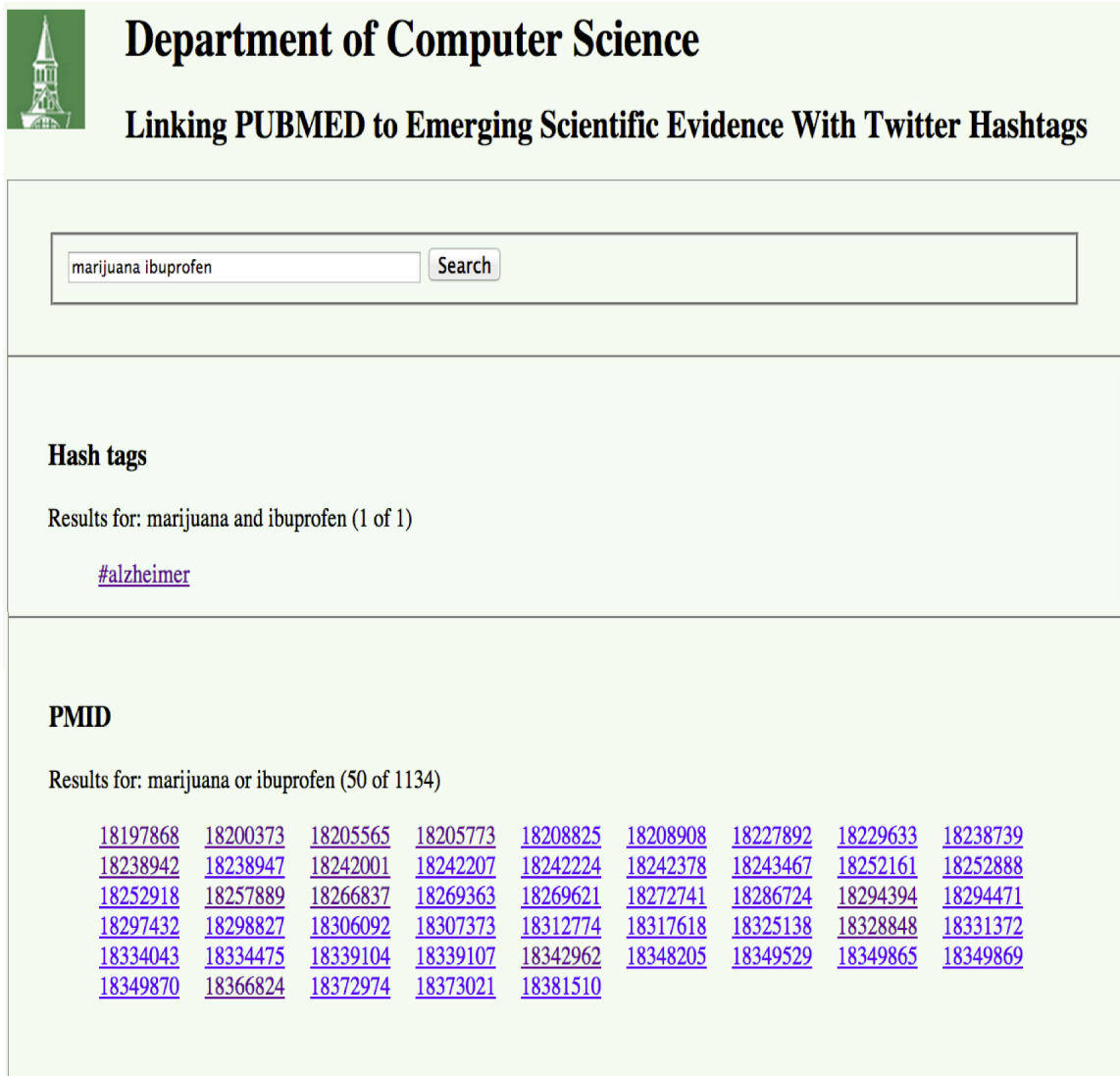
## 4.6 Conclusions

In this chapter , we have presented an association analysis proof of concept experiments to demonstrate the importance of hashtags in terms of the type of knowledge they carry. We two different type of associations: (1) constructed from keywords that co-occurred in the same tweets (1) associations between keywords and hashtags also appeared in the same tweet. From a proof of concept standpoint, we have found that emergent hashtags (i.e., entirely new that emerged unexpected) have different types and carry within much knowledge from the outside world. Just from 25 tweets we learned various types of knowledge fluctuated among geolocations, date/time, acronyms, events, synonyms and many more. Such wealth can reveal much more of the hidden details. Therefore, using hashtags in computational analysis is bound to produce interesting discoveries. We also found various types of hashtags that do not have any syntactic similarity with the MeSH terms used. The lesson learned from assessing hashtags made it clear that there is much to be discovered beneath. In Section 4.4 we showed a much more comprehensive study using millions of tweet to enable such discoveries.

Following the preliminary section of the chapter , we have introduced a systematic method that detects patterns from hashtag-based networks. We presented a network mining algorithm, which we



Figure 4.6: PubMed Mesh-Hashtag Linking Web Application



**Department of Computer Science**

**Linking PUBMED to Emerging Scientific Evidence With Twitter Hashtags**

marijuana ibuprofen Search

**Hash tags**

Results for: marijuana and ibuprofen (1 of 1)

[#alzheimer](#)

**PMD**

Results for: marijuana or ibuprofen (50 of 1134)

<a href="#">18197868</a>	<a href="#">18200373</a>	<a href="#">18205565</a>	<a href="#">18205773</a>	<a href="#">18208825</a>	<a href="#">18208908</a>	<a href="#">18227892</a>	<a href="#">18229633</a>	<a href="#">18238739</a>
<a href="#">18238942</a>	<a href="#">18238947</a>	<a href="#">18242001</a>	<a href="#">18242207</a>	<a href="#">18242224</a>	<a href="#">18242378</a>	<a href="#">18243467</a>	<a href="#">18252161</a>	<a href="#">18252888</a>
<a href="#">18252918</a>	<a href="#">18257889</a>	<a href="#">18266837</a>	<a href="#">18269363</a>	<a href="#">18269621</a>	<a href="#">18272741</a>	<a href="#">18286724</a>	<a href="#">18294394</a>	<a href="#">18294471</a>
<a href="#">18297432</a>	<a href="#">18298827</a>	<a href="#">18306092</a>	<a href="#">18307373</a>	<a href="#">18312774</a>	<a href="#">18317618</a>	<a href="#">18325138</a>	<a href="#">18328848</a>	<a href="#">18331372</a>
<a href="#">18334043</a>	<a href="#">18334475</a>	<a href="#">18339104</a>	<a href="#">18339107</a>	<a href="#">18342962</a>	<a href="#">18348205</a>	<a href="#">18349529</a>	<a href="#">18349865</a>	<a href="#">18349869</a>
<a href="#">18349870</a>	<a href="#">18366824</a>	<a href="#">18372974</a>	<a href="#">18373021</a>	<a href="#">18381510</a>				

call *HashnetMiner*, that searches for heuristics (open triads and entirely disconnected nodes in the network of word associations). The algorithm uses the keyword association network as a lookup mechanism and a mean of identifying the patters in the Hashtag-based network. Through this exploratory process, we have opened new door for analyzing Hashtags, designed a new algorithm that has proven promising in terms of exposing hidden paths among concepts. Showing the knowledge

that can be discovered by linking MeSH and Hashtags can reveal much details about current public health events happening on social media that publications alone can not reveal. Hashtagging publications revives the static PubMed Digital Library and makes it lively. Especially, if the hashtags are periodically updated to reflect the current events and trends. For instance, a scientist should be able to see why a hashtag is trending and what MeSH terms are associated with it.


A pattern such as #alzheimer linking marijuana and Ibuprofen was a great motivation to further investigate what can be mined. Upon reverse-engineering this pattern and examining the original tweets we found that there maybe a correlation between Ibuprofen via the #alzheimer hashtag. This was surprising because it was a conflict with what Gorsky stated: “*Marijuana test: no ibuprofen interference.*” and published in the Science Magazine in 1989 (Gorsky, 1988). However, a recent study has revived this link between and suggests further discussions. Hashtagging articles related to marijuana and Ibuprofen is indeed a much needed feature that MeSH alone can no longer achieve due to the static aspect of it. Figure 4.7 shows some of the traffic available when this news emerged. By incorporating hashtags to existing Medline articles, scientists can stay up to date with current research finding, and also investigate new hypothesis based on the new evidence that hashtags can bring in. Clearly, the knowledge gained from such surprising discoveries is beyond imagination.

The patterns detected from *HashnetMiner* only derives information on an anecdotal basis and cannot given at face value a medical conclusion without substantive testing and challenging. This by no means diminishes the value of information extrapolated from this database. Each pattern has a value on its own basis. Each pattern represents a premise of its own and deserves a comprehensive examination to see if any deserves a medical outcome test. No observation of this database can stand on its own as a basis of facts or theory, the only relation between the database and an individual Twitter users is to promote clinical discussion. The sole purpose of our database initiative is to identify correlations between two or possible three drugs that humans use for various reasons. *HashnetMiner* have demonstrated how hashtags can connect nuggets of information and synthesize new knowledge. In the drug interaction part of this chapter, Hashtags has proved colorful

of various types (diseases: #aids, symptoms:#heartburn, drugs names: #prednisone, substance synonyms:#hemp). The findings of our algorithms have demonstrated that hashtags are a great source of discoveries when linked to concepts of a particular domain. Mapping and linking such discoveries to a digital library such as PubMed can be very promising and offers a new method of bringing anecdotal information into clinical thinking. This opens the doors for many possibilities and yields a new way to extrapolate information using means of data mining and knowledge discovery algorithms.


The future directions of this chapter many. A compelling possibility is from a computational standpoint, which requires further development of the *HashnetMiner* algorithm. This may include testing the need of incorporating new patterns in addition to the currently existing ones. This will require a much deeper experimentation and analysis to confirm the usefulness of each pattern to be added. Another possibility is on the application front by extending the scope of the current study beyond the very limited number drugs used in previous experiments presented in this chapter . Yet another possibility, is to design the benchmark network from existing drug-drug interaction databases as a ground truth. This also includes mining PubMed Medline articles for all known interactions. The scope can also extend to study the adverse side effects of known interactions or newly discovered ones. This may generate hypotheses and confirms theories about symptoms that occur from intaking two drugs at the same time and by the same person. Enhancing the web client to directly link to PubMed using the LinkOut service (NLM, 2005b) may yield a much more useful outcome. LinkOut is a web service which enables full-text and supplemental information to be directly linked from the Medline abstracts. This can make dissemination of knowledge discovered by *HashnetMiner* a powerful tool to clinicians, scientists and all players in the Biomedical Web Science community.

Figure 4.7: Marijuana, Ibuprofen and Alzheimer on Twitter


Results for **marijuana #alzheimer ibuprofen** 

[Top](#) / [All](#) / [People you follow](#)


---

 **Pure Essence Labs** @PureEssenceLabs 26 Nov  
RT @latimeshealth: To turn **marijuana** into a brain-calming drug 4 **#Alzheimer's**, study suggest taking it w **ibuprofen**. [lat.ms/1i2MjvE](http://lat.ms/1i2MjvE)  
[View summary](#) [Reply](#) [Retweet](#) [Favorite](#) [More](#)


---

 **420times** @4twentytimes 26 Nov  
Photo: growswitch: **#Marijuana** with a side of **ibuprofen**: Rx for buzz-kill and **#Alzheimer's** Disease View Post [tumblr.co/ZbjAhx-UaL2X](http://tumblr.co/ZbjAhx-UaL2X)  
[View photo](#) [Reply](#) [Retweet](#) [Favorite](#) [More](#)

---

 **#Alzheimer** @AlzTuz 25 Nov  
**#Memory** Loss From **Marijuana** Blocked By **Ibuprofen**; Drug Duo May Halt **#Alzheimer's** Progression [medicaldaily.com/memory-loss-ma...](http://medicaldaily.com/memory-loss-ma...) [#alzheimers](#)  
Expand [Reply](#) [Retweet](#) [Favorite](#) [More](#)

---

 **Mildred Rivera** @MildredRiveraD 24 Nov  
**Marijuana** with a side of **ibuprofen**: Buzz-killing Rx for **#Alzheimer's**? [latimes.com/science/scienc...](http://latimes.com/science/scienc...)  
[View summary](#) [Reply](#) [Retweet](#) [Favorite](#) [More](#)

## 4.7 Acknowledgments

The authors of this chapter would like to thank Dr. Tamás Nepusz and the iGraph team for their help with visualization and valuable discussions. The authors also would like to thank Dr. Luis Rocha of Indiana University, Center of complex Networks and Systems Research for the valuable feedback on linking drugs to MeSH. A special thanks to Robert Erickson of University of Vermont, Department of Computer Science for his help developing the Web Interface.

## Chapter 5

# Conclusions and Future Work

### 5.1 Conclusions

In this dissertation we pose a few research questions which are related to Twitter hashtag knowledge discovery: How feasible is it to develop models that are entirely based on hashtags to aid in discovering useful knowledge on Twitter? Can these models be general enough to apply across domains? Is it possible to answer questions such as: Can we produce the base components for new recipes? Can we come up with new Ice Cream flavors? New coffee flavors? Can we organize items in grocery stores more efficiently? Can we learn about adverse side effects that occur as a result of taking two drugs together, beyond what we discovered in this thesis? Can we learn about unreported drug side effects? Can we recommend new cars based on sentiments expressed in tweets? We can make new book recommendations based on books people tweet about on Twitter? Can we understand diseases and their causes better? Can we learn how outbreaks erupt and how they spread? More importantly, can we come up with a formal model that addresses all these questions and many more?

In the preliminary analysis, we demonstrated that Hashtags are useful via means of association rule mining. We have presented a notion of ground truth to compare a workbench model against a hashtag-based model. Using very small datasets, we learn better rules with higher confidence when hashtags are incorporated than when they are ignored. This finding helped to highlight the useful

gain of hashtags. This finding, however, didn't discriminate which rules are significant and which rules should be pruned out.

In Chapter 2, we presented a Twitter rule-based system that uses association rule mining to recruit Twitter users for two different medical studies: (1) Smoking cessation using nicotine patches, and (2) A world-wide study of health and environment conducted over the web. The system used an expert system shell in the backend to decide when to recruit a subject based on the hashtags included in the tweet. This system is dynamic and ever-evolving since the knowledge acquisition engine is constantly providing newly computed hashtag-based association rules.

In Chapter 3, we performed an exploratory analysis to test the hypothesis of whether hashtags can make Big Data environments (Twitter in this case) smaller. We designed two network models: (1) a keyword-keyword network model to use as a benchmark where hashtags are not included, and (2) keyword-hashtag network model, which we call (K-H networks). We replicated the Milgram experiments, especially of Backstrom et al., which was done on a Big Data environment. Using millions of tweets, we tested the connectivity of a finite set of given keywords (domain specific keywords (cars, sentiments, drugs, smoking substances)). We posed the question: Does the shrinking world phenomenon, presented by Backstrom et al., apply to our two network models? In this study we presented an exploratory analysis of large-scale K-H networks generated from Twitter. We used two different measures: (1) the degrees of separation, and (2) the eccentricity of keyword vertices, a well known centrality measure. Our analysis shows that K-H networks conform to the phenomenon of the shrinking world presented by Backstrom et al. Specifically, it reveals that the number of vertices between any two keywords that were not originally connected in the K-K networks is exactly three. The eccentricity of every keyword in the K-H networks is four. The previous experiments shed light on the importance path patterns mining using our network models.

In Chapter 4, we posed the question: Can Twitter K-H networks provide clinicians with a powerful tool to extrapolate patterns that may lead to the development of new medical therapies and/or drugs? In this paper, we present a novel algorithm which we call *HashnetMiner*. The algorithm

is designed to discover new patterns and operate on both K-K and K-H networks. We applied the algorithm to the biomedical field, concerning Drug-Drug Interaction (DDI) and marijuana-drug interaction. Concepts are selected from widely accessible databases (e.g., Medical Subject Heading [MeSH] descriptors, and Drugs.com). The algorithm has led to promising discoveries which could suggest further avenues of future research.

This research contributes a novel network model (K-H network models) and a pattern mining algorithm which operates on those K-H networks. We provide a case study proving this new model discovered information about marijuana-drug interaction. We have then linked the discoveries back to medical related publications in order to facilitate further investigation.

## 5.2 Future Work

This thesis has opened many doors and generated more questions that we believe are candidates for future research.

In Chapter 2, we discussed a rule-based system to recruit Twitter users to participate in a nicotine patch study. This chapter provided a preliminary analysis based on the traffic that was generated and the number of impressions that were made. However, no further analysis was provided. This is due to the fact that the study is designed to release the various statistics at the end of year 2015. Another possible direction is to address the complexity of programming in CLIPS/JESS. The learning curve of mastering such programming languages is steep. It is particularly challenging to design an industrial strength system without extensive training. We believe we will interface the system with a pure Java implementation to make it possible for expert programmers to advance the design and development of the current system and naturally integrate knowledge acquisition tools (e.g., Weka and SPMF).

In Chapter 3, we simulated the K-K and K-H networks using the Erdős and Rényi's (ER) random network models to better understand the structure of our networks. However, the ER models cannot adequately model real-world networks. For example, the K-H networks demonstrated that

particular vertices are more central than others (e.g., marijuana and happy). Such a behavior cannot be modeled by the ER networks due to the fixed probability that dictates how many edges can be connected to each vertex. To address this limitation, other random network models (e.g., Watts and Strogatz) (Watts & Strogatz, 1998) must be used to verify whether the findings of this chapter are reproducible using purely random networks. Additionally, the experiments of chapter 3 use a finite set of input keywords from various domains (e.g., drugs, sentiments, car models, etc). The findings of this chapter are specific to the datasets and keywords used. We believe that it is worthwhile to perform the same type of experiments using all words that appear in tweets, not only a given set of keywords. This requires a very large dataset of billions of tweets gathered over a long period of time. This type of analysis will be able to account for words related to holidays and events (e.g., Christmas, Thanksgiving) and show a much more realistic results. Removing noise words (a.k.a. stop-words) and performing stemming of words are expected preprocessing steps for a more concrete analysis and more significant results.

In Chapter 4, we presented the *HashnetMiner* algorithm and experiments for discovering drug-drug and marijuana-drug interaction. Medical marijuana has become legal in 20 American states and the District of Colombia (DC). There is an ongoing political movement to legalize medical marijuana in the U.S.A. and in other countries. This new reality brings new challenges, such as understanding which drugs have adverse side effects when prescribed simultaneously with medical marijuana. With the rapid movement of marijuana legalization. Identifying evidence from social collective behavior, as captured by millions of tweets, may lead to further investigations of actual marijuana-drug interactions (e.g., such as pharmacokinetic in vitro, invico, and clinical studies). Developing a database of such interactions and making it publicly available would be useful to identifying potential adverse reaction cases. We aim to extend the scope of the work of this dissertation and perform a very large-scale analysis using billions of tweets gathered over many years. The new experiments will also make use of K-K networks, which we plan to construct using positive controls from DDI databases. Such controls will contain known interactions among drugs (E.g.



tums and protonix for heartburn). This aids in identifying the negative controls and mark them as candidates for unknown interactions. This chapter presented various example of negative controls (marijuana and clarinex) and (marijuana and ibuprofen) interactions.

Throughout this dissertation, we provide ideas for further research that may contribute to both the theoretical and practical aspects of knowledge discovery using our network models and algorithm.

### **5.3 Acknowledgments**

The authors would like to thank the Center for Complex Networks and Systems Research, School of Informatics Computing, Indiana University for providing several years of garden-hose tweets.

# References

- Abel, F., Marenzi, I., Nejd, W., & Zerr, S. (2009, August). LearnWeb2.0: Resource Sharing in Social Media. In R. Vuorikari, H. Drachsler, N. Manouselis, & R. Koper (Eds.), *Workshop on social information retrieval for technology-enhanced learning (sirtel'09) at the international conference on web-based learning (icwl '09), aachen, germany* (Vol. Vol-535). CEUR-WS.org. Retrieved from [http://ceur-ws.org/Vol-535/Abel\\_SIRTEL09.pdf](http://ceur-ws.org/Vol-535/Abel_SIRTEL09.pdf)
- Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th international conference on very large data bases* (pp. 487–499). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Retrieved from <http://dl.acm.org/citation.cfm?id=645920.672836>
- Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2012). Four degrees of separation. In *Proceedings of the 3rd annual acm web science conference* (pp. 33–42). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/2380718.2380723> doi: 10.1145/2380718.2380723
- Berlingerio, M., Bonchi, F., Bringmann, B., & Gionis, A. (2009, September). Mining graph evolution rules. In W. L. Buntine, M. Grobelnik, D. Mladenic, & J. Shawe-Taylor (Eds.), *Machine learning and knowledge discovery in databases*, (pp. 115–130). Springer. Retrieved from <https://lirias.kuleuven.be/handle/123456789/247409> doi: 10.1007/978-3-642-04180-8
- Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2011). The socialbot network: When

- bots socialize for fame and money. In *Proceedings of the 27th annual computer security applications conference* (pp. 93–102). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/2076732.2076746> doi: 10.1145/2076732.2076746
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7), 107–117.
- CareerBuilder. (1995). <http://www.careerbuilder.com>.
- Carter, S., Tsagkias, E., & Weerkamp, W. (2011, 06/2011). Twitter hashtags: Joint translation and clustering. In *Web science 2011*. Koblenz: ACM.
- Castells, M., & Cardoso, G. (2006). *The network society: From knowledge to policy*. Center for Transatlantic Relations, Paul H. Nitze School of Advanced International Studies, Johns Hopkins University. Retrieved from <http://books.google.com/books?id=pZJgNvEMnyMC>
- Chen, E. S., & Sarkar, I. N. (2010, June). Meshing molecular sequences and clinical trials: A feasibility study. *J. of Biomedical Informatics*, 43(3), 442–450. Retrieved from <http://dx.doi.org/10.1016/j.jbi.2009.10.003> doi: 10.1016/j.jbi.2009.10.003
- Clegg, B. (2012). *Gravity: How the weakest force in the universe shaped our lives*. St. Martin's Press. Retrieved from <http://books.google.com/books?id=W1V1L0TCpqc>
- Com, D. D. (2001). *www.drugs.com*. Retrieved from <http://www.drugs.com>
- Conner, C. (2008). *A people's history of science: Miners, midwives, and "low mechanics"*. Paw Prints. Retrieved from <http://books.google.com/books?id=eWeoPwAACAAJ>
- Conover, M., Gonçalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011). Predicting the political alignment of twitter users. In *Proceedings of 3rd ieee conference on social computing (socialcom)*. Retrieved from [http://cnets.indiana.edu/wp-content/uploads/conover\\_prediction\\_socialcom\\_pdfexpress\\_ok\\_version.pdf](http://cnets.indiana.edu/wp-content/uploads/conover_prediction_socialcom_pdfexpress_ok_version.pdf)
- Costantino, C. M., Gupta, A., Yewdall, A. W., Dale, B. M., Devi, L. A., & Chen, B. K. (2012,

- 03). Cannabinoid receptor 2-mediated attenuation of cxcr4-tropic hiv infection in primary cd4+ t cells. *PLoS ONE*, 7(3), e33961. Retrieved from <http://dx.doi.org/10.1371/journal.pone.0033961> doi: 10.1371/journal.pone.0033961
- Craigslist. (1995). <http://www.craigslist.com>.
- Csardi, G., & Nepusz, T. (2006). The igraph Software Package for Complex Network Research. *InterJournal, Complex Systems*, 1695. Retrieved from <http://igraph.sf.net>
- D'Amico, N. (2009). Pinpointing gravity. *Science*, 323(5919), 1299-1300. Retrieved from <http://www.sciencemag.org/content/323/5919/1299.short> doi: 10.1126/science.1170936
- Deyne, S. D., & Storms, G. (2008). Word associations: Network and semantic properties. *Behavior Research Methods*, 40(1), 213-231. Retrieved from <http://brm.psychonomic-journals.org/content/40/1/213.abstract>
- Dodds, P., & Danforth, C. (2009). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*. (<http://dx.doi.org/10.1007/s10902-009-9150-9>) doi: 10.1007/s10902-009-9150-9
- Dodds, P. S., Muhamad, R., & Watts, D. J. (2003). An experimental study of search in global social networks. *Science*, 301, 827-829. Retrieved from <http://dx.doi.org/10.1126/science.1081058>
- Erdős, P., & Rényi, A. (1959). On random graphs. I. *Publ. Math. Debrecen*, 6, 290-297.
- Facebook. (2004). <http://www.facebook.com>. Retrieved from <http://www.kbb.com>
- Ferreira, P. G., & Starkmann, G. (2009). Einstein's Theory of Gravity and the Problem of Missing Mass. *Science*, 326, 812-815. doi: 10.1126/science.1172245
- Fortunato, S., Boguñá, M., Flammini, A., & Menczer, F. (2005). How to make the top ten: Approximating pagerank from in-degree. *CoRR*, *abs/cs/0511016*.
- Fortunato, S., Boguñá, M., Flammini, A., & Menczer, F. (2007). On local estimations of pagerank: A mean field approach. *Internet Mathematics*, 4(2), 245-266.
- Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L. (2010). Weka-

- A Machine Learning Workbench for Data Mining. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 1269–1277). Boston, MA: Springer US. Retrieved from [http://dx.doi.org/10.1007/978-0-387-09823-4\\_66](http://dx.doi.org/10.1007/978-0-387-09823-4_66) doi: 10.1007/978-0-387-09823-4-66
- Freund, Y., & Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the second european conference on computational learning theory* (pp. 23–37). London, UK, UK: Springer-Verlag. Retrieved from <http://dl.acm.org/citation.cfm?id=646943.712093>
- Ganjisaffar, Y. (2012). *Open source web crawler for java*. Retrieved from <http://code.google.com/p/crawler4j/>
- Good, B., Tennis, J., & Wilkinson, M. (2009). Social tagging in the life sciences: characterizing a new metadata resource for bioinformatics. *BMC Bioinformatics*, *10*(1), 1-17. Retrieved from <http://dx.doi.org/10.1186/1471-2105-10-313> doi: 10.1186/1471-2105-10-313
- Google. (1995). <http://www.google.com>.
- Gorsky, J. (1988). Marijuana test: no ibuprofen interference. *Science*, *241*, 888. doi: 10.1126/science.3043663
- Gross, J. L., & Yellen, J. (2005). *Graph theory and its applications, second edition (discrete mathematics and its applications)*. Chapman & Hall/CRC.
- Gupta, N., Das, A., Pandey, S., & Narayanan, V. K. (2012). Factoring past exposure in display advertising targeting. In *Proceedings of the 18th acm sigkdd international conference on knowledge discovery and data mining* (pp. 1204–1212). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/2339530.2339719> doi: 10.1145/2339530.2339719
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009, November). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, *11*(1), 10–18. Retrieved from <http://doi.acm.org/10.1145/1656274.1656278> doi: 10.1145/1656274.1656278
- Hanaki, N., Peterhansl, A., Dodds, P. S., & Watts, D. J. (2007, July). Cooperation in evolving social

- networks. *Manage. Sci.*, 53(7), 1036–1050. Retrieved from <http://dx.doi.org/10.1287/mnsc.1060.0625> doi: 10.1287/mnsc.1060.0625
- Hand, D. J., & Yu, K. (2001). Idiot's bayesnot so stupid after all? *International Statistical Review*, 69(3), 385–398. Retrieved from <http://dx.doi.org/10.1111/j.1751-5823.2001.tb00465.x> doi: 10.1111/j.1751-5823.2001.tb00465.x
- Hanneman, R. A., & Riddle, M. (2005). Introduction to social network methods [Computer software manual]. Riverside, CA. Retrieved from <http://www.faculty.ucr.edu/~hanneman/>
- Hastie, T., & Tibshirani, R. (1996, June). Discriminant adaptive nearest neighbor classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(6), 607–616. Retrieved from <http://dx.doi.org/10.1109/34.506411> doi: 10.1109/34.506411
- Hill, E. F. (2003). *Jess in action: Java rule-based systems*. Greenwich, CT, USA: Manning Publications Co.
- Hougardy, S. (2010). The floyd-warshall algorithm on graphs with negative cycles. *Inf. Process. Lett.*, 110(8-9), 279-281. Retrieved from <http://dblp.uni-trier.de/db/journals/ipl/ip1110.html#Hougardy10>
- Incorporation, B. (2013). *Url shortening and bookmarking services*. Retrieved from <http://bitly.com/>
- Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1), 59 - 68. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0007681309001232> doi: <http://dx.doi.org/10.1016/j.bushor.2009.09.003>
- KBB. (1996). *Kelly Blue Book: www.kbb.com*. Retrieved from <http://www.kbb.com>
- Keyhole.co. (2007). *Hashtag analytics @ONLINE*. Retrieved from <http://www.keyhole.co/>
- Kleinberg, J. M. (1999, September). Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5), 604–632. Retrieved from <http://doi.acm.org/10.1145/324133.324140> doi: 10.1145/324133.324140

- Lehmann, J., Gonçalves, B., Ramasco, J. J., & Cattuto, C. (2012). Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on world wide web* (pp. 251–260). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/2187836.2187871> doi: 10.1145/2187836.2187871
- Linehan, D. (2011). *Spaceshipone: An illustrated history*. MBI Publishing Company. Retrieved from <http://books.google.com/books?id=3v9H1J8ZVtAC>
- LinkedIn. (2003). <http://www.linkedin.com>.
- MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In L. M. L. Cam & J. Neyman (Eds.), *Proc. of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 1, p. 281-297). University of California Press.
- Marvel, S. A., Martin, T., Doering, C. R., Lusseau, D., & Newman, M. E. J. (2013). The small-world effect is a modern phenomenon. *CoRR*, *abs/1310.2636*. Retrieved from <http://dblp.uni-trier.de/db/journals/corr/corr1310.html#MarvelMDLN13>
- Mazman, S., & Usluel, Y. (2010). Modeling educational usage of facebook. *Computers Education*, *55*(2), 444 - 453. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0360131510000424> doi: <http://dx.doi.org/10.1016/j.compedu.2010.02.008>
- McCarthy, J. J., McLeod, H. L., & Ginsburg, G. S. (2013, June 12). Genomic Medicine: A Decade of Successes, Challenges, and Opportunities. *Science Translational Medicine*, *5*(189), 189sr4. Retrieved from <http://dx.doi.org/10.1126/scitranslmed.3005785> doi: 10.1126/scitranslmed.3005785
- McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley Series in Probability and Statistics.
- Meng, X., Wei, F., Liu, X., Zhou, M., Li, S., & Wang, H. (2012). Entity-centric topic-oriented opinion summarization in twitter. In *Proceedings of the 18th acm sigkdd international conference on knowledge discovery and data mining* (pp. 379–387). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/2339530.2339592> doi: 10.1145/

2339530.2339592

- Monster. (1999). *http://www.monster.com*. Retrieved from <http://www.monster.com>
- Newman, M. E. J. (2000). Models of the small world. *J. Stat. Phys*, 819–841.
- Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review*, 45(2), 167–256. Retrieved from <http://scitation.aip.org/getabs/servlet/GetabsServlet?prog=normal&id=SIREAD000045000002000167000001&idtype=cvips&gifs=yes>
- NLM. (2005a, May). *Medline/pubmed data element (field) descriptions @ONLINE*. Retrieved from <http://www.nlm.nih.gov/bsd/mms/medlineelements.html#rn>
- NLM. (2005b, July). *Viewing linkout resources in ncbi database records: The linkout display*. Retrieved from [http://www.ncbi.nlm.nih.gov/books/NBK3810/#using.Viewing\\_LinkOut\\_Reso](http://www.ncbi.nlm.nih.gov/books/NBK3810/#using.Viewing_LinkOut_Reso)
- Noh, J. D., & Rieger, H. (2004). Random walks on complex networks. *Physical Review Letters*, 92, 118701. Retrieved from doi:10.1103/PhysRevLett.92.118701
- Palla, G., Farkas, I. J., Pollner, P., Derenyi, I., & Vicsek, T. (2007). Directed network modules. *New Journal of Physics*, 9(6), 186. doi: 10.1088/1367-2630/9/6/186
- Pavlyshenko, B. (2013). Forecasting of events by tweet data mining. *CoRR*, abs/1310.3499. Retrieved from <http://dblp.uni-trier.de/db/journals/corr/corr1310.html#Pavlyshenko13>
- Perakis, F., Mattheakis, M., & Tsironis, G. P. (2014). *Small-world networks of optical fiber lattices*. Retrieved from <http://arxiv.org/abs/1401.2321>
- Pring, C. (2012). *99 new social media stats for 2012*. Retrieved from <http://thesocialskinny.com/99-new-social-media-stats-for-2012/>
- Priss, U., & Old, L. J. (2007, July). Bilingual word association networks. In U. Priss, S. Polovina, & R. Hill (Eds.), *Proceedings of the 15th international conference on conceptual structures (iccs 2007)* (Vol. 4604, p. 310-320). Berlin, Heidelberg: Springer-Verlag.



- Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Rapoport, A. (1953). Spread of information through a population with socio-structural bias: I. assumption of transitivity. *The bulletin of mathematical biophysics*, 15(4), 523-533. Retrieved from <http://dx.doi.org/10.1007/BF02476440> doi: 10.1007/BF02476440
- Ravikumar, S., Talamadupula, K., Balakrishnan, R., & Kambhampati, S. (2013). Raprop: Ranking tweets by exploiting the tweet/user/web ecosystem and inter-tweet agreement. In *Aaai (late-breaking developments)* (Vol. WS-13-17). AAAI. Retrieved from <http://dblp.uni-trier.de/db/conf/aaai/late2013.html#RavikumarTBK13>
- Ritter, A., Mausam, Etzioni, O., & Clark, S. (2012). Open domain event extraction from twitter. In *Proceedings of the 18th acm sigkdd international conference on knowledge discovery and data mining* (pp. 1104–1112). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/2339530.2339704> doi: 10.1145/2339530.2339704
- Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on world wide web* (pp. 695–704). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/1963405.1963503> doi: 10.1145/1963405.1963503
- Strogatz, S. H. (2001, March 08). Exploring complex networks. *Nature*, 410(6825), 268–276. Retrieved from <http://dx.doi.org/10.1038/35065725> doi: 10.1038/35065725
- Tasneem, A., Aberle, L., Ananth, H., Chakraborty, S., Chiswell, K., McCourt, B. J., & Pietrobon, R. (2012, 03). The database for aggregate analysis of clinicaltrials.gov (aact) and subsequent regrouping by clinical specialty. *PLoS ONE*, 7(3), e33677. Retrieved from <http://dx.doi.org/10.1371/journal.pone.0033677> doi: 10.1371/journal.pone.0033677
- Ting, R. M. H., & Bailey, J. (2006). Mining minimal contrast subgraph patterns. In J. Ghosh, D. Lambert, D. B. Skillicorn, & J. Srivastava (Eds.), *Sdm*. SIAM. Retrieved from <http://>

dblp.uni-trier.de/db/conf/sdm/sdm2006.html#TingB06

- Travers, J., & Milgram, S. (1969, December). An experimental study of the small world problem. *Sociometry*, 32(4), 425–443.
- Twitter.com. (2006). *The twitter rest api*. Retrieved from <https://dev.twitter.com/docs/api>
- Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc.
- Wang, X., Tokarchuk, L., Cuadrado, F., & Poslad, S. (2013). Exploiting hashtags for adaptive microblog crawling. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 311–315). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/2492517.2492624> doi: 10.1145/2492517.2492624
- Watts, D. J. (2004). *Six degrees: The science of a connected age*. W. W. Norton & Company. Paperback. Retrieved from <http://www.amazon.ca/exec/obidos/redirect?tag=citeulike04-20{\&}path=ASIN/0393325423>
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 409–10.
- Wu, X., & Kumar, V. (2009). *The top ten algorithms in data mining* (1st ed.). Chapman & Hall/CRC.
- Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., ... Steinberg, D. (2007, December). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1), 1–37. Retrieved from <http://dx.doi.org/10.1007/s10115-007-0114-2> doi: 10.1007/s10115-007-0114-2
- Yamaguchi, Y., Takahashi, T., Amagasa, T., & Kitagawa, H. (2010). Turank: twitter user ranking based on user-tweet graph analysis. In *Proceedings of the 11th international conference on web information systems engineering* (pp. 240–253). Berlin, Heidelberg: Springer-Verlag. Retrieved from <http://dl.acm.org/citation.cfm?id=1991336.1991364>
- Yamamoto, Y. (2007). *Java library for the twitter api*. Retrieved from <http://www.twitter4j>

.org/

- Yan, X., & Han, J. (2002). gspan: Graph-based substructure pattern mining. In *Icdm* (p. 721-724). IEEE Computer Society. Retrieved from <http://dblp.uni-trier.de/db/conf/icdm/icdm2002.html#YanH02>
- Yang, M., Lee, J.-T., Lee, S.-W., & Rim, H.-C. (2012). Finding interesting posts in twitter based on retweet graph analysis. In *Proceedings of the 35th international acm sigir conference on research and development in information retrieval* (pp. 1073–1074). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/2348283.2348475> doi: 10.1145/2348283.2348475
- Yang, X., Ghoting, A., Ruan, Y., & Parthasarathy, S. (2012). A framework for summarizing and analyzing twitter feeds. In *Proceedings of the 18th acm sigkdd international conference on knowledge discovery and data mining* (pp. 370–378). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/2339530.2339591> doi: 10.1145/2339530.2339591
- Yu, K., Ding, W., Simovici, D. A., & Wu, X. (2012). Mining emerging patterns by streaming feature selection. In *Proceedings of the 18th acm sigkdd international conference on knowledge discovery and data mining* (pp. 60–68). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/2339530.2339544> doi: 10.1145/2339530.2339544
- Zhang, B., Wang, J., & Zhang, L. (2013). A tweet-centric algorithm for news ranking. In *Icdcs workshops* (p. 190-195). IEEE. Retrieved from <http://dblp.uni-trier.de/db/conf/icdcs/icdcs2013.html#ZhangWZ13>
- Zhang, J.-L., Zhang, J.-W., & Xu, L. (2010). Fpg-grow: A graph based pattern grow algorithm for application level io pattern mining. In *Proceedings of the 2010 ieee fifth international conference on networking, architecture, and storage* (pp. 311–316). Washington, DC, USA: IEEE Computer Society. Retrieved from <http://dx.doi.org/10.1109/NAS.2010.23> doi: 10.1109/NAS.2010.23