DETECTING POLICY VIOLATORS IN ONLINE SOCIAL COMMUNITY
AN EXTENDED BAYESIAN BELIEF NETWORK APPROACH

A Thesis in
Computer Science
by
Shuo Huang

© 2012 Shuo Huang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2012
The thesis of Shuo Huang was reviewed and approved* by the following:

Suncun Zhu  
Associate Professor  
Thesis Advisor

Anna Cinzia Squicciarini  
Assistant Professor  
Thesis Advisor

John Yen  
Director, Full Professor

Raj Acharya  
Director, Full Professor  
Head of the Computer Science and Engineering Department

*Signatures are on file in the Graduate School
ABSTRACT

In this thesis, I have implemented a solution for detecting policy violators in online social communities. Given the increasing number of users and traffic in online social services, e.g., forums, it is difficult for administrators to manually oversee the activities. My solution is designed to resolve such problems.

To achieve the goal, this thesis implemented a risk warning system, using Bayesian networks (BN). BN is, firstly, a directed acyclic graph. Each node in the BN represents a hypothesis for user to have certain attributes. The arcs describe causal relationships between nodes. In this thesis, BN is designed using naïve-based classifier. In other words, it is assumed that all hypotheses are independent to each other. Secondly, the BN is also a statistical model. The data collected represents behavioral features about a user. These features, after processed by input nodes, become hypothesize of user. Input nodes are parent nodes of intermediate nodes. Each intermediate node represents an intermediate attribute of the user. Intermediate nodes are parents of core model nodes. Core model nodes model intent, opportunity and capability of the monitored user. These core model nodes are parents of the result node. For the result of BN, the result node produces a conditional probability. This value indicates a binary outcome of whether the user is malicious.

Our solution includes a number of key techniques for handling data mining and processing. For example, Page Rank and Degree Centrality are implemented to monitor the popularity of the user. Sentimental Analysis, Topic Mining and Mutual Information are used for detecting the Authenticity and Relevance of user generated content.

For test and evaluation, this thesis commandeers real world data of a top forum. The tests are designed to include fifty users with more than ten thousand posts per user as sample. To evaluate the performance of the BN, true positive and true negative ratios are carefully counted.
against human moderator. A true positive happens when BN successfully catches an abusive post. A true negative occurs when the post is malicious but the test result indicating otherwise. The test results show very high ratio of true positives and true negatives comparing to human moderator.
# TABLE OF CONTENTS

LIST OF FIGURES ........................................................................................................ vi

LIST OF TABLES .......................................................................................................... vii

ACKNOWLEDGEMENTS .............................................................................................. viii

Chapter 1 Introduction ................................................................................................. 1

Chapter 2 Related Works .............................................................................................. 5

Chapter 3 Preliminary Notions ..................................................................................... 8

  3.1 Bayesian Network ............................................................................................... 8
  3.2 Page Rank ........................................................................................................... 12
  3.3 Mutual Information ............................................................................................ 14
  3.4 Degree Centrality .............................................................................................. 15

Chapter 4 Core Solution .............................................................................................. 17

  4.1 Detailed Analysis of Input Nodes ....................................................................... 18
  4.2 Detailed Analysis of Intermediate Nodes .......................................................... 25
  4.3 Detailed Analysis of Core Model Nodes ............................................................. 27
  4.4 Detailed Analysis of the Result Node ................................................................. 28

Chapter 5 Extended Solution ...................................................................................... 29

  5.1 Detailed Analysis of Input Nodes ....................................................................... 33
  5.2 Detailed Analysis of Intermediate Nodes .......................................................... 37
  5.3 Detailed Analysis of Core Model Nodes ............................................................. 38

Chapter 6 Algorithms .................................................................................................. 41

Chapter 7 Design and Implementation ....................................................................... 52

  7.1 Backbone System ............................................................................................... 52
  7.2 Plugin Tool .......................................................................................................... 61

Chapter 8 Test and Result ........................................................................................... 63

Chapter 9 Future Works ............................................................................................... 72

Chapter 10 Conclusion ................................................................................................. 73

Reference ..................................................................................................................... 74

Appendix Tables ......................................................................................................... 85
LIST OF FIGURES

Figure 3-1 Sample Bayesian Network Configuration .......................................................... 8
Figure 3-2 Sample Page Rank Configuration A ................................................................. 13
Figure 3-3 Sample Page Rank Configuration B ................................................................. 14
Figure 3-4 Graphical Representation of Mutual Information ............................................. 15
Figure 4-1 Bayesian Network Configuration of Core Solution ......................................... 17
Figure 5-1 Extended Sub-graph of Node Intent ................................................................. 30
Figure 5-2 Extended Sub-graph of Node Capability ......................................................... 31
Figure 5-3 Bayesian Network Configuration of Extended Solution ................................. 32
Figure 5-4 Components of the Extended Solution System .............................................. 40
Figure 7-1 IO Package Class Diagram ............................................................................. 53
Figure 7-2 Key Package Diagram ..................................................................................... 54
Figure 7-3 Basic Data Structure Package Diagram .......................................................... 56
Figure 7-4 Exception Package Class Diagram ................................................................. 57
Figure 7-5 Data Management Package Diagram ............................................................. 58
Figure 7-6 Node Package Diagram .................................................................................. 59
Figure 7-7 Word Processor Package Class Diagram ......................................................... 59
Figure 7-8 Procedure Page Class Diagram ..................................................................... 60
Figure 7-9 Screenshot of Plugin Tool ............................................................................... 61
LIST OF TABLES

Table 3-1 CPT of Hypothesis Is Sober ................................................................. 9
Table 3-2 CPT of Hypothesis Take Key ................................................................. 9
Table 3-3 CPT of Hypothesis Put On Belt .............................................................. 9
Table 3-4 CPT of Hypothesis Take License ......................................................... 9
Table 3-5 CPT of Hypothesis Drive Legally ......................................................... 10
Table 4-1 Input Node Output Values and Conditions ......................................... 23
Table 5-1 Input Node Output Values and Conditions ......................................... 36
Table 8-1 Quantified Test Results Comparison for One User ......................... 68
Table 8-2 Quantified Test Results Comparison for Average for Fifty Users ....... 70
Table B-1 CPT for Character .............................................................................. 85
Table B-2 CPT for Cost ..................................................................................... 86
Table B-3 CPT for Benefit .................................................................................. 87
Table B-4 CPT for Authenticity ......................................................................... 89
Table B-5 CPT for Activity ................................................................................. 90
Table B-6 CPT for Intent .................................................................................... 92
Table B-7 CPT for Opportunity ......................................................................... 93
Table B-8 CPT for Capability ............................................................................ 94
Table B-9 CPT for Threat .................................................................................. 98
Table B-10 CPT for Influence ......................................................................... 99
Table B-11 CPT for Content ............................................................................ 101
Table B-12 CPT for Centrality ......................................................................... 102
ACKNOWLEDGEMENTS

In this acknowledgement, I want to thank Dr. Sencun Zhu, Dr. John Yen and Dr. Anna Cinzia Squicciarini for all their instructions, guidance and help on my academic studies, researches, direction of career and many other aspects during my stay at PSU. Without their directing, I will not be able to achieve what I am today.

I want to thank Dr. Anna Cinzia Squicciarini, who allowed me work in her laboratory. Through this experience, I have gain tremendous knowledge on security, computer programming, cloud computing, data mining and many other aspects of computer science and information technology. She also advised me on this thesis, for which, I cannot be more grateful. She is very patient, tolerant with her students and yet strict to academic and research disciplines and standards. I would say it is my greatest honor to have you as my thesis advisor.

I want to thank Dr. John Yen for giving me the opportunity for working for him as a Research Assistant. It is this indispensable experience that allows me to have a deep insight on cloud computing, parallel data processing and data mining. And it is also because of this experience, I landed job positions that I have dreamed of.

I want to thank Dr. Sencun Zhu. Although I did not take any of his courses, he has provided me guidance and help beyond academic knowledge. I have learnt so much from him. He is a teacher to me and also a friend.

In the end, I want to thank Jing Huang and Xiaoying Fan, who have given me everything a person could. I would not be able to come to PSU without your support during this period and all of the times that came before. I am forever in your debt. And also Beichen Tang, thank you for your company, help and patience. I love you all.
Chapter I

Introduction

Online social network services (SNS) have been popular for almost a decade now. According to Tech Crunch report [2], Over 1.4 billion users worldwide are active on SNSs. SNSs are effective for helping users maintain their social contacts.

A number of full-fledged SNS services lead market. Facebook is the most influential name. It also has the best market share [5]. Facebook was launched in February 2004 [6]. Facebook is the first SNS that supports a well-designed relation network with unique components, e.g., the “wall”.

As an add-on to Facebook, Zynga [8] developed a number of hyper-addictive games, such as Farmville and Cityville [9]. These games allow existing users of Facebook to build farms and cities in virtual world. You can choose to help your friends build their virtual property or steal their efforts when they are not online.

Besides Facebook, there are similar services, e.g., Google+, LinkedIn, Foursquare and Quora. Google+ [10] is a service launched by Google to compete with Facebook. Google+ now has more than 170-million users [10]. Different from Facebook, Google+ allows users to categorize their friends into circles. The “stream”, which is the Google version of “Wall”, allows users to share text, picture and videos only to the specific circles of friends. This feature has greatly enhanced privacy options as users may share only certain information with people he chose. Other than this feature, Google+ is the first SNS to support in-browser video chat.

Apart from Facebook and Google+, LinkedIn, Foursquare and Quora focus on niche markets. LinkedIn is a career centric social network [11] with 100-million users.
Before modern SNS, most of the social network websites are built as a bulletin board system. To date, a large number of such services still exist. Tianya Club [13] is the 12th most visited site in China and 72nd worldwide. Forums are usually focused on a topic. For example, the xda-developer forum [14] contains topics like mobile phone application and operating system development. Like other forums, xda-developer has many sub-sections. Each section has different topics. Users, when posting contents, submit contents related to the topic in that subsection. To administer the site, normally there is at least one super administrator that oversees all sections. For each subsection, at least one dedicated administrator is assigned. This thesis focuses on experimenting on platforms like this, although the methodology and algorithms can be applied to all forms of SNS.

One of the major problems of forums is lack of consistent controls of users’ actions within the sites. Policy violators often ignore the sites’ terms of use. Their behavior includes grieving, trolling, flaming, harassment, threats trolling, multiple accounts, shared accounts, advertising, plagiarism and etc.

Cases of data misuse happen every day [15] [16] [17]. One example in these reports describes a case of abusing SNS and forums using short URLs for phishing. Spammers post comments with short URLs to lure users to click on malicious content. Once clicked, the victim will automatically install Trojan horse software on his computer. Currently, the enforcement for supervising the site over site-policy in user-contributed sites is largely operated manually by moderators. These moderators are often dedicated and long-running members with good site reputation, who are devoted to patrol and oversee their site and take action against members displaying deviant behavior. However, given the size and activity rate of some forums, it is difficult for moderators to oversee all the users’ posts. Although some automatic tools exist to detect vandalism and bots [18] [19] [17], none of them takes site-policy and user generated content into account.
The objective of this thesis is to resolve such problems, by means of an automated warning system. The thesis introduces a user-centered model using Bayesian Belief Network (BN) [19]. BNs have already been used in different scenarios to model uncertainty, and have been proven effective for real-time security analysis.

In the thesis, two models of the solutions are exemplified, core solution and extended solution.

Core solution is originated from the paper [21] and is a part of the effort for this thesis. The core solution employs BN that takes input of real-time features of the behavior of users. After going through a series of decision nodes, the network calculates the threat probability of the user.

Extended solution introduces a number of new technologies to produce better results. A tailored version of the Page Rank algorithm is implemented in the network as an input node to rank user’s influence.[20]. The more influential a user is, the more potential harmfulness he can impede onto the forum. Another major improvement in this thesis is to monitor the actual content of users. The original Bayesian network only takes the sentimental value and bad word count of user content into account. The current version of the network also detects the mutual information between user posts and threads. If a user abuses the network by posting malicious content or advertisement, the actual content of such posts are likely to have low relevance with the content posted by others. Lastly, Degree Centrality [21] is introduced to examine the post and reply ratio of users. For advertisers, the number of posts is normally much higher than the ones that have been replied.

To our current knowledge, this thesis discusses the first user-centered model for policy enforcement in online sites.

This thesis has eleven chapters. Chapter 4 introduces the core solution from the original paper [21]. Chapter 5 explains extended solution in detail. Chapter 6 exemplifies the key
algorithms implemented in this thesis. Chapter 7 explains the design and implementation of the program and demo case in detail. Chapter 8 shows test designs and results.
Chapter 2
Related Works

There are a number of noticeable publications related to Human Social Behavior in recent years. The virtual world of SNS is a mirror to the real world. Often times, user behaviors of the latter are found in its counterpart. The behaviors, e.g., posts, relationships and personal influences are effective traces for analyzing user character. Gao et al. [23] study relationships between social network structure and knowledge sharing. Their work focuses on discovering the reasons behind different knowledge sharing rate by social network topologies. Wan et al. along with other scholars study on how social network structure evolves [24] [25] [26] [27]. Such findings are related to identification of rules for structural evolution in social network. These “rules” take the number of strong and weak ties into account. Mohsen J. et al. [28] define metrics for analyzing social network, e.g., centrality, maximum flow, strong and weak ties, cliques, n-clans and k-plexes. To analyze user dynamics in social networking sites, user behavior must be quantified with metrics that yield statistical or mathematical means [29] [30]. E.g., using the metrics defined by Mohsen J and et al, users can be categorized in different types. Using Centrality, Ding et al. [31] analyze social features in bulletin board systems. Kang W. et al. [32] conduct research on social networks during emergency times. They analyze the information sharing features during the Japanese earthquake and nuclear event.

The research efforts of social behavior have uncovered features of spam related to SNS. Pedram H. et al. [33] define the latest forms of spam. According to the publication, Pedram H. et al introduce a wide variety of spam. Some notorious spam including phishing, robot messaging and advertising on social platforms are well explained in this paper. These new spam thrive after
the booming of web 2.0 and online social networks. Apart from the definitions of spam, the influences are exemplified by Man Q. et al [34]. Man Q. et al introduce the motivations, operation approaches and effects of spam. According to the work, the major motivations of spam are economic benefits. Influences and consequence of spam to the society can be extremely negative. Such influences can cause mistrust in the banking systems and government by the public. Besides spam, there are other kinds of abusive posts discussed in Malicious and Spam Posts in Online Social Networks by Saeed A. et al. [35]. Saeed A. et al. analyze data from Facebook. They discover that besides well-designed and organized spam, malicious posts generated by legitimate users also count a large fraction of abusive posts. The topic spam detection is a well-developed field with large number of publication related [36] [37] [38] [39]. The previous efforts on building detection system are mostly focused on isolated metrics and analysis on individual posts. Such techniques do not integrate the notion of beliefs and intention that are instead considered in this thesis. Hongyu G. et al. [40] create similar method described in [36] [37] with modeling and clustering “Wall” posts of Facebook and URL analyze for the detection of spam campaigns. Some other research efforts focus on detecting spam in comment systems, such as the analysis of YouTube [29] and others [41]. These papers introduce individual metrics for analyzing comments system similar to the input nodes of this thesis and Squicciarini et al. [19]. The methodology introduced in this thesis and [19] employs Bayesian Belief Network (BN). BN is superior than the works related to spam detections because BN not only has a more comprehensive set of inputs, but is also a multi-parent and multi-child graph that governed by the Bayesian probability rules. Thus, it produces more accurate predictions that those works. Similar to BN, Facebook has developed an immune system based on the deterministic automaton [42]. The system contains only fixed number of states. It takes a number of input nodes, such as context of the posts, total
number of posts, friends, interests and etc. The automaton also contains states and intermediate nodes. However, the decision of the choices on the path from one node to another is not determined by probability. For instance, if user’s total number of posts passed the preset limit, the immune system will definitely go from the current state to a designated state. Such a system is, first of all, designed and optimized only for Facebook. Second, the decision on the user can sometimes be too deterministic. Different from this system, the design in this thesis allows multiple possible paths upon the same input value at one point and hence allows more flexibility, possibility and depth on the decision making process.

This thesis implements Page Rank as input nodes for BN. There are limited works on Page Rank related to social network analysis. Shaojie Q. et al. [43] describes an innovative page rank approach based on similarities of contents. Michalis V. et al. [44] introduce method to combine Page Rank and Markov Chain Models to predict web page ranking. There are also a number of efforts made for building customized Page Rank engines [45] [46] [47] [48]. This thesis has a customized implementation specially optimized bulletin board system.

Another component in this thesis is content analysis. Numerous research efforts have been devoted to related fields. One of the most popular techniques is top mining. K-means is a common model for extracting topics out of text [49], Xiang W. et al. [50] and Erim et al. [51] describe approaches on top of K-means to search for topics out of asynchronous text streams, e.g., text messaging thread.
Chapter 3

Preliminary Notions

3.1 Bayesian Network

Bayesian Belief Network (BN) is the core statistical model implemented in this thesis. The name BN comes from the probabilistic inference process of the network, by which it follows Bayes’ rules. In some contexts, the BN is also known as Belief Network. A BN is also a probabilistic graphical model. Within this model, the nodes represent random variables, and the arcs describe the causal relationship between the parent and child node. Different from Markov Random Fields, BN is directed graphical model.

BN contains nodes, arcs and Conditional Probability Distributions (CPDs). The nodes are hypothesizes indicating the chances for events to happen. The Conditional Probability

Figure 3-1 Sample Bayesian Network Configuration
Distribution describes the probabilities for hypotheses to be true. CPD is represented as Conditional Probability Tables (CPT).

Table 3-1 CPT of Hypothesis Is Sober

<table>
<thead>
<tr>
<th>P(IS = True)</th>
<th>P(IS = False)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Table 3-2 CPT of Hypothesis Take Key

<table>
<thead>
<tr>
<th>IS</th>
<th>P(TK=False)</th>
<th>P(TK = True)</th>
</tr>
</thead>
<tbody>
<tr>
<td>False</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>True</td>
<td>0.9</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 3-3 CPT of Hypothesis Put On Belt

<table>
<thead>
<tr>
<th>IS</th>
<th>P(POB = False)</th>
<th>P(POB = True)</th>
</tr>
</thead>
<tbody>
<tr>
<td>False</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>True</td>
<td>0.2</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Table 3-4 CPT of Hypothesis Take License

<table>
<thead>
<tr>
<th>IS</th>
<th>P(TL = False)</th>
<th>P(TL = True)</th>
</tr>
</thead>
<tbody>
<tr>
<td>False</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>TK</td>
<td>POB</td>
<td>TL</td>
</tr>
<tr>
<td>----</td>
<td>-----</td>
<td>----</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

Table 3-5 CPT of Hypothesis Drive Legally

Figure 3-1 shows an example BN. Table 3-1 through table 3-5 list the truth tables for each node in the BN. All nodes have only two possible values. The values are denoted by T (true) and F (false). Each node represents a hypothesis of event. This BN provides a structural logic flow of legal driving. For instance, if the driver is sober and he has driver license, car key and he put on the safety belt, then he has a very high chance of driving home legally. However, he might forget his license or forget his key. He will need to hot wire his car and drive home without a license. Then he is not driving home legally. The node on the bottom represents the event “drive legally” (DL = true) has multiple possible causes, Pr(DL=true | TK=true, POB=true, TL=true, Pr(DL=true | TK=true, POB=true, TL=false), Pr(DL=true | TK=true, POB=false, TL=true) and
Pr(DL=true | TK=false, POB=true, TL=true) The simplest condition independent relationship in a Bayesian network is that a node is independent of its ancestors given its parents, where the parent-child relationship is according to the topological ordering of the nodes.

\[
P(IS, TK, POB, TL, DL) = P(IS) \times P(TK|IS) \times P(POB|TK, IS) \times P(TL|POB, TK, IS) \times P(DL|TL, POB, TK, IS)
\]

By using conditional independence relationships, we can rewrite this as

\[
P(IS, TK, POB, TL, DL) = P(IS) \times P(TK|IS) \times P(POB|IS) \times P(TL|IS) \times P(DL|TL, POB, TK)
\]

The most common task to solve, using Bayesian networks, is probabilistic inference. For example, consider the drive legally network, and suppose the driver did drove legally. There are four possible causes for this. For the events of have key, have license and put on belt, at least two of these events must be true. Which is more likely? We can use Bayes' rule to compute the posterior probability of each explanation (where 0=false and 1=true).

\[
Pr(TK = 1, POB = 1, TL = 1|DL = 1) = \frac{\sum_{TL_1, IS_1} Pr(TL = TL_1, TK = 1, POB = 1, IS = IS_1, DL = 1)}{Pr(DL = 1)} = \frac{0.07425}{0.36655} = 0.2027
\]

Equation 3-3
This is called “explaining away”. In statistics, this is known as Berkson’s paradox, or “selection bias”.

3.2 Page Rank

PageRank is a technique to efficiently rank the importance of a node in a graph. When a node has an edge to another node, it is effectively casting a vote to that node. By counting the number of votes the node has, the importance of this node can be calculated using Page Rank. With the Page Rank values, we can rank the importance of all the nodes in the network. In the context of this thesis, users can be seen as nodes and their posts are edges. Hence, Page Rank is used for ranking the importance of users.

To compare Page Rank to Closeness Centrality, Closeness Centrality measures the inverse of farness node. Farness is the sum of distance between the examined nodes to all other nodes. Thus, if there are two nodes, the nodes link to each other, this is the scenario for Closeness Centrality to achieve maximum value. If the distance between nodes are the same, as the number of nodes increases, the value of Closeness Centrality decreases. This is not a good metrics to model user popularity in a social network. If the user has a large number of repliers, then he is less popular than those with fewer repliers. This does not make sense. Closeness Centrality does not take the popularity of other nodes into account. For example, if two users have the same
number of repliers. But if the repliers of one user never reply to others but the repliers of the other user replies to a large number of other users, then the popularity of these two users is quite different. However, Closeness Centrality cannot detect this difference. Page Rank on the other hand can resolve all these disadvantages Closeness Centrality has.

PageRank is denoted as PR for the following equations.

\[
PR(A) = (1 - d) + d \frac{PR(t_1)}{C(t_1)} + \cdots + d \frac{PR(t_n)}{C(t_n)}
\]

Equation 3-4

In Eq. 3-5, \( t_n \) represents the nodes. \( PR(t_n) \) is the Page Rank value for node \( t_n \). For simplicity, \( PR(t_n) \) is usually initialized to \( 1/n \). \( C(t_n) \) is the number of inbound edges from \( t_n \) to node A and \( d \) is the damping factor. The damping factor is normally set to a constant of 0.85.

For clarity, we provide some simple examples.

![Figure 3-2 Sample Page Rank Configuration A](image)

In Figure 3-2, the example contains 3 nodes and 6 edges. Each node has two outbound links towards another node as well as two inbound links. In this case,

\[
PR(A) = 0.25 + 0.85 * \frac{PR(B)}{2} + 0.85 * \frac{PR(C)}{2}
\]

\[
PR(B) = 0.25 + 0.85 * \frac{PR(A)}{2} + 0.85 * \frac{PR(C)}{2}
\]
PR(C) = 0.25 + 0.85 * PR(B) / 2 + 0.85 * PR(A) / 2

Then PR(A) = PR(B) = PR(C) = 1.67184

Figure 3-3 Sample Page Rank Configuration B
The second example in figure 3-3 has three nodes, but only node A has an outbound link toward node B.

PR(B) = 0.25 + 0.85 * PR(A) / 1 = 1
PR(A) = 0.25
PR(C) = 0.25

3.3 Mutual Information

As its name, Mutual Information is a technique in information theory for finding the probability of shared knowledge from pieces of independent information. Mutual information describes the joint intersection of two sets of random variables. Intuitively, it is represented by the probability of the common knowledge that two pieces of independent information shares. It is appropriate to use this technique to find similarities between two pieces of texts.
Figure 3-4 exemplifies an example of mutual information. $H(X \mid Y)$ and $X(Y \mid X)$ represent the marginal entropy of random variable $X$ and $Y$. $I(X \mid Y)$ is the mutual information.

Hence the mutual information can be computed as the following equation,

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right)$$

$P(X, Y)$ is the joint probability of $x$ and $y$.

3.4 Degree Centrality

Degree Centrality is often used to measure the density of the connections associated with a node. In a directed graph, the degree centrality can be divided into in-degree centrality and out-degree centrality. The in-degree centrality measures the connection density of the edges linked inwards to the node while out-degree centrality counts the density of the edges pointing outwards from the node. The degree centrality is a good metric for measuring the activeness of the traffic in
and out of a node. In this thesis, the application of this metric can decide the activeness of a user by measuring the density of his post read by others and the density of the replies he received. For abusive users, the ratio of the in-degree and out-degree centrality is often extremely unbalanced. For example, advertisers and spam posters often has a much higher density of out-degree centrality than in-degree centrality.

The degree of a node is calculated using the equations below. The function $C_D(v)$ is the degree centrality function for node $v$. Function $\text{deg}(v)$ measures the degree of the connected sub-graph linked to node $v$.

$$C_D(v) = \text{deg}(v) \quad \text{Equation 3-6}$$

$$\text{deg}(v) = |E| + |V| \quad \text{Equation 3-7}$$

Degree centrality for a graph is calculated by:

$$C_D(G) = \frac{\sum_{i=1}^{\mid V \mid} [C_D(v \ast) - C_D(v_i)]}{H} \quad \text{Equation 3-8}$$

Where

$$H = \sum_{j=1}^{\mid V \mid} C_D(y \ast) - C_D(y_j) \quad \text{Equation 3-9}$$
Chapter 4
Core Solution

The model of the core solution introduced in [20] is constructed based on Bayesian Belief Network (BN) concept. The structure of the model is shown in Figure 4-1. The model contains twenty-one nodes, which include ten input nodes, five intermediate nodes, three core nodes and one result node.

The BN is created to model the behaviors of users. For normal behaviors, we define the behaviors of the majority to be normal. Using this base line, there are two extremes of user behaviors, the extremely good and extremely bad. The goal of the solution is to capture as much user in the bad section as possible.
The following example demonstrates an instance of using the Bayesian. At the beginning, the input nodes gather user behaviors as inputs. Using these inputs, the input nodes produces hypothesizes for the user to have a series of input attributes. The hypothetical attributes are passed to the child nodes, the intermediate nodes. The intermediate nodes employ the noisy or logic in the CPT to produce intermediate hypothesizes. The output values of these nodes are discrete states. For example, the node Authentic Behavior outputs values one, two or three. The value one is mapped to above average, two means neutral and three represents below average. Following such a pattern, the final hypothesis is produced by the outputs from Intent, Opportunity and Capability. At the end of this round of BN operation, the threat node produces a probability indicating whether the user is predicted as a policy violator. In the design of core solution, naïve-based classifier is implemented to form the assumption that all the attributes of the user are independent to each other. The following paragraphs exemplify the details of the operation processed by each node.

4.1 Detailed Analysis of Input Nodes

In the following, we discuss the detailed input requirements, detailed functionalities and output values of the input nodes.

1. Post Behavior,

Post behavior (PB) measures the number of posts submitted to the forum within a period of time by a specific topic by the examined user. PB needs only the userid as input. Using the userid, PB calls the designated Java application to retrieve all the
posts associated with userid. Then PB compares the PB value with the average value among all users in the forum. Eq. 4-1 describes how the value is actually computed. In the equation, Th denotes thread and p represents post.

\[
PB^{Th_i}_T (u) = \sum_{v|p \in Th_i, i \leq cl_w} 1
\]

Equation 4-1

PB produces integer value as output. Output values and their meanings are shown in table 4-1.

2. Change in Post Behavior,

Change in Post Behavior (CPB) measures the value change in the value of PB as time passes by. CPB needs userid as input. PB output for the current timestamp and the timestamp from half of the time period is calculated. CPB measures the level of change in two PB values,

\[
CPB^{Th_i}_T (u) = \frac{PB^{Th_i}_T (u) - PB^{Th_i}_T (u)}{PB^{Th_i}_T (u)}
\]

Equation 4-2

Output values and their meanings are shown in table 4-1.

3. Authentic Behavior,

Authentic Behavior (AB) measures the ratio of conforming posts (CP) against the total posts submitted by the user. Two criteria are required to test before determining whether a post belongs to the conformed category. Firstly the node examines if there are any cursed words in the post. Secondly, the node checks the sentiment value of
the post. Both parts count half of the value in the function CP. AB requires userid as input. The result is then used to compare with the mean value of all users in the forum. An integer value is produced as result. Eq. 4-3 shows the way the value is calculated. In this equation, TP denotes the function that calculates the total number of posts by the examined user and CP represents the number of conforming posts.

\[
AB^{Th}_{i} (u) = \frac{CP^{Th}_{i} (u)}{TP^{Th}_{i} (u)}
\]

Output values and their meanings are shown in table 4-1.

4. Change in Authentic Behavior,

Change in Authentic Behavior (CAB) measures the change of the AB value through time. It requires userid as input. Similar to the CPB, the values compared are the AB values of current timestamp and the half time.

\[
CAB^{Th}_{i} (u) = \frac{AB^{Th}_{i} (u) - AB^{Th}_{i, prev} (u)}{AB^{Th}_{i, prev} (u)}
\]

Output values and their meanings are shown in table 4-1.

5. Contribution Behavior,

The Contribution Behavior (CB) measures the overall impact of user by monitoring the contents of the user’s posts. CB requires userid as input. Equation 4-5 describes
the details of how to compute value CB. In the equation, NCP denotes the total number of non-conforming posts.

\[
CB_i^{Th_k}(u) = \frac{CP_i^{Th_k}(u) - NCP_i^{Th_k}(u)}{TP_i^{Th_k}(u)}
\]

Equation 4-5

Output values and their meanings are shown in table 4-1.

6. Change in Contribution Behavior,

Change in Contribution Behavior (CCB) indicates the change in the output value from the node CB. Similar to the CPB, the values compared are the CB values of current timestamp and the half time.

\[
CCB_i^{Th_k}(u) = \frac{CB_i^{Th_k}(u) - CB_i^{Th_k,prev}(u)}{CB_i^{Th_k,prev}(u)}
\]

Equation 4-6

Output values and their meanings are shown in table 4-1.

7. Target Posts,

Target Posts (TP) measures the total number of posts the forum currently have.

8. Target Users,

Target Users (TU) measures the total number of users in the forum.

9. Previous Sanction,
Previous Sanction (PS) measures the number of points fined by moderators when any misbehaves of the user is detected. For the test forum in this thesis, the site has a point-based punishment-system. The point fined can be an integer from one to ten. When the user has accumulated points over a certain limit, his privileges may be deprived, e.g., submitting posts. To the worst, he can be exiled from the forum indefinitely. The node PS compares the value of currently examined user to the mean value of all users and produces an integer value output. In Eq. 4-7, function p denotes the function for retrieving total number of points fined for the user.

\[ PS(u) = \sum p(u) \]  

Equation 4-7

Output values and their meanings are shown in table 4-1.

10. Forum Attention,

Forum Attention measures the number of actions taken by the moderator to the user, e.g., move of post. The node Forum Attention compares the value of currently examined user to the mean of all users and results in an integer value output.

\[ FA(u) = \sum a(u) \]  

Equation 4-8

Output values and their meanings are shown in table 4-1.

11. Access,

Access measures the access privilege of the user about whether the site is designed for the particular system. For example, some sites have a set of access codes for user.
The levels of privileges are represented by integer values. Value zero normally indicates the user is registered but not activated. One means the user is registered and activated. Two suggests the user is the administrator of the section. Three indicates the user has a higher-level administrator position. Within this thesis, unified value one is set for all tested users.

12. Resources,

Node Resources represents the resources that can be accessed by user. For example, some sites provide different resources for different users, e.g., according to the amount of time the user stay in the forum or the contribution made to the forum. Unified value 1 is set for all tested users.

<table>
<thead>
<tr>
<th>Table 4-1 Input Node Output Values and Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>Post Behavior</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Change in Post Behavior</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Category</td>
</tr>
<tr>
<td>----------------------------</td>
</tr>
<tr>
<td><strong>Authentic Behavior</strong></td>
</tr>
<tr>
<td><strong>Change in Authentic Behavior</strong></td>
</tr>
<tr>
<td><strong>Contribution Behavior</strong></td>
</tr>
<tr>
<td><strong>Change in Contribution Behavior</strong></td>
</tr>
<tr>
<td><strong>Previous Sanction</strong></td>
</tr>
<tr>
<td><strong>Forum Attention</strong></td>
</tr>
<tr>
<td><strong>Access</strong></td>
</tr>
</tbody>
</table>
4.2 Detailed Analysis of Intermediate Nodes

Next, we discuss detailed input requirements, functionalities and outputs of the intermediate nodes.

1. Character,

Node Character defines the character of the user. It requires two inputs, CB and CCB. There are three output states: spontaneous, contained and persistent. Table B-1 shows the combinations of the input values and their corresponding results. For example, if CB = 3 and CCB = 3, the Character of this user has 57% chance of being persistent, 29% chance of being spontaneous and 14% chance of being contained.

2. Costs,

Node Costs determines the cost for the forum to oversee the actions of user. It takes three inputs, Character, PS and FA. There are three output states, which are major cost, some cost and no cost at all. For the value major cost, this result indicates that the user posts a large number of posts and comments with a high frequency. Within which a large number of his posts have caught the attention of the moderator. Thus, the moderator will need to frequently examine his posts and take actions. For the value no cost, either the user has very few posts and comments or most of his posts
do not show traces abnormal activity. Table B-2 shows the combinations of the input values and their corresponding results.

3. Benefits,

Node Benefits measures the degree of benefit the user has brought to the forum. It takes three inputs: Character, TP and TU. The three output states categorized the user effort into three levels: one. The user has major benefit to the forum; two the user has some benefits to the forum. And three, the user has no benefits to the forum. Intuitively, the more positive the Character the user is and the more posts and users are the higher benefits the user can provide the forum. Table B-3 shows the combinations of the input values and their corresponding results.

4. Authenticity,

Node Authenticity measures the level of authenticity of user’s posts. It requires two inputs, AB and CAB. It produces two possible outputs, inauthentic and authentic. Table B-4 shows the combinations of the input values and their corresponding results.

5. Activity,

Node Activity measures the activeness of user. It requires two inputs, PB and the CPB. It produces two possible output states, active and inactive. This output value is determined by the activeness of this user and the average activeness of all users.
Table B-5 shows the combinations of the input values and their corresponding results.

**4.3 Detailed Analysis of Core Model Nodes**

In the following, we discuss the detailed input requirements, functionalities and output options of the core model nodes.

1. **Intent,**

   Node Intent measures user intention. Intent requires two inputs, Authenticity and Activity. Node Authenticity provides traces of how the user is posting abusive posts while Node Activity measures how the user is abusing the forum through the frequency of his posts. Table B-6 shows the combinations of the input values and their corresponding results.

2. **Opportunity,**

   Node Opportunity examines the opportunity that the user succeeded in becoming a threat to the forum. Two input nodes connect to Opportunity, Cost and Benefit. Node Cost reflects the cost of moderator to oversee user. As the size of a forum grows, the cost of overseeing the entire forum also magnifies. Benefit measures the contribution the user makes to the community. When the cost for overseeing the user is high and he produces negative benefits to the forum, the user have great opportunity to pollute
the forum. Table B-7 shows the combinations of the input values and their corresponding results.

3. Capability,

Node Capability measures the capability user has in terms of affecting the forum. The capability is determined by Access and Resources. Access indicates the areas of the forum the user may abuse. Resources represent the ways the user may abuse the forum. Table B-8 shows the combinations of the input values and their corresponding results.

4.4 Detailed Analysis of the Result Node

Finally, node Result produces final result for the current round of BN execution. Through hundreds of rounds of experiment, a threshold is trained to classify result value. If result value is larger than threshold, the BN determines that user is malicious. Consequently, he is a threat to the forum. Table B-9 shows the combinations of the input values and their corresponding results.
Chapter 5

Extended Solution

The original BN discussed in the previous section is effective for analyzing user history. However, two important factors are omitted, social contexts and message content analysis. Although the original BN examines posts on cursed words and sentimental values, there are no nodes analyzes the content by their nature.

To address this gap, we introduce an extended model. Six new input nodes and three intermediate nodes are added to the original BN. The new input nodes are Social Influence, Change in Social Influence, Content Relevance, Change in Content Relevance, Degree Centrality and Change in Degree Centrality. The new intermediate nodes are node Influence, Content and Centrality.

The first group of nodes added to the original BN is for monitoring the contextual features of the post. This group has three nodes, two of which are input nodes, Content Relevance and Change in Content Relevance. Node Content Relevance measures similarities in topics between posts and threads. Node Change in Content Relevance checks the level of change in value of Content Relevance. These input nodes have implementations of two important technologies, topic mining and mutual information. The nodes first use topic mining retrieve topics in texts and compares similarity with threads using mutual information. The other node is Content node. This is an intermediate node. It gathers outputs of Content Relevance and Change in Content Relevance and produces a result. This result tells if the overall post contents is relevant to the threads.

This group of nodes is attached to the Intent node. Now the intent node has three inputs, Authenticity, Activity and Content. Figure 5-1 shows the nodes and edges in the Intent branch.
The second group of nodes attached to BN measures influence of the user. It has three nodes, Social Influence, Change in Social Influence and Influence. Inspired by Page Rank [12], which is originally used to rank web pages, influence of users can be measures by same means. In Google’s version of Page Rank, the popularity of a website is measured by counting the number of wen pages that have a link to this website. The more web pages linked to a particular web page, the higher the latter is. Similarly, in social network, we can consider users as nodes and number replies to his posts as edges. Intuitively, in a bulletin board system, if user posts are replied by most users in the forum, then he is considered influential. Node Social Influence ranks users using this approach. Change in Social Influence measures the level of changes in the values of Social Influence. The third node in this group is the node Influence. This is an intermediate node, which takes the outputs of Social Influence and Changes in Social Influence to produce a result. This result indicates the level of influence the user has over the forum.
The third group of nodes added measures the ratio of in-degree and out-degree centrality of the user. This value is a ratio between inward edges and outward edges in the social graph. In the forum, we can consider the users as nodes, and their posts and comments to others are outward edges, and the replies they received are inward edges. The ratio of the inward and outward edges reveals two features of the user. If the ratio is very high this means, the user has far more replies than he posts. In other words, the user is very influential, since a large number of people are interested in his posts. On the other hand, if the ratio is very low, the user posts significantly more than reply. In such a case the user is likely to be a spammer. This group contains three nodes, Degree Centrality, Change in Degree Centrality and Centrality. The node Degree Centrality measures degree centrality. Change in Degree Centrality checks the level of changes in the node Degree Centrality. The third node in the group is Centrality. Centrality takes the output values of Degree Centrality and Change in Degree Centrality and produces an output by the CPT table associated with.
Figure 5-2 shows the extended branch of Capacity node. Now the capacity node takes four inputs, Resources, Access, Influence and Centrality.
5.1 Detailed Analysis of Input Nodes

In the following, we discuss the detailed input requirements, functionalities and output values of the input nodes.

1. Social Influence,

Node Social Influence (SI) calculates the Page Rank value of the user given the current timestamp. This node takes userid as input. It will gather all the postids from the posts of user. Then the userids of the repliers and the number of posts of these repliers are acquired. Later, these results are fed into the Eq. 5-1 to calculate Page Rank value.

\[
PR(u_0) = \frac{PR(u_1)}{RP(u_1,u_0)} + \frac{PR(u_2)}{RP(u_2,u_0)} + \ldots + \frac{PR(u_n)}{RP(u_n,u_0)}
\]

Equation 5-1

In Eq. 5-1, \(PR(u_0)\) here represents the Page Rank value of a user, and \(RP(u_i,u_0)\) denotes the number of post replies of \(u_i\) replied to \(u_0\). This node compares the Page Rank value of the current user to the average Page Rank value of all users in the forum. For output, it produces an integer value.

Output values and their meanings are shown in table 5-1

1. Change in Social Influence,

Node Change in Social Influence (CSI) measures the change in the value (SI) of a given user. It takes userid as input. CSI calculates the Page Rank values of this user for the current timestamp and half the time period before. For the output, it produces an integer value indicating the level of change happened in the aspect of SI.
\[ CPR(u_0) = \frac{PR_{\text{curr}}(u_0) - PR_{\text{prev}}(u_0)}{PR_{\text{prev}}(u_0)} \]

Equation 5-2

Output values and their meanings are shown in table 5-1

2. Content Relevance

Node Content Relevance (CR) measures the relevance of the content of the posts the user submitted against each of the threads these posts belong to. Eq. 5-3 and Eq. 5-4 describe the detail of how this value is calculated. In these equations, \( P_i \) means the post with id equals to \( i \) and \( Th_i \) represents the thread that \( p_i \) belongs to.

\[ CR(u_0) = \frac{\sum_{Th_i} R(p_i, Th_i)}{\sum_i} \]

Equation 5-3

\[ R(p_i, Th_i) = I(p_i; Th_i) = \sum_{y \in Th_i} \sum_{x \in P_i} p(x, y) \log\left( \frac{p(x, y)}{p(x)p(y)} \right) \]

Equation 5-4

\( R(p_i, Th_i) \) is the relevance function of each post regarding to their thread. After computed the value of CR, the node compares this value to the average CR of all the users in the forum. As the output, it produces an integer value. Output values and their meanings are shown in table 5-1

3. Change in Content Relevance

Change in Content Relevance (CCR) measures the change of value of CR. It takes userid as input. CCR calculates the CR values of this user for the current timestamp
and half the time period before. For the output, it produces an integer value indicating
the level of change happened in the aspect of CR.

$$CCR(u_0) = \frac{CR_{\text{cur}}(u_0) - CR_{\text{prev}}(u_0)}{CR_{\text{prev}}(u_0)}$$

Equation 5-5

Output values and their meanings are shown in table 5-1

4. Degree Centrality

Degree Centrality (DC) calculates the in and out degree ratio of the user. DC takes
userid as input and computes the in/out degree ratio as shown in Eq. 5-6 and Chapter
3. DC compares this value with the average DC value of all the users in the forum. It
produces an integer value as output.

$$DC(u_0) = \frac{ID(u_0)}{OD(u_0)}$$

Equation 5-6

In Eq. 5-6, ID() is the function for calculating the in-degree centrality given the
userid $u_0$ and OD() is for out-degree centrality. Output values and their meanings are
shown in table 5-1

5. Change in Degree Centrality

Change in Degree Centrality (CDC) measures the change in DC. It takes userid as
input. CDC calculates the DC values of this user for the current timestamp and half
the time period before. For the output, it produces an integer value indicating the
level of change happened in the aspect of DC.
Equation 5-7

\[
CDC(u_0) = \frac{DC_{curr}(u_0) - DC_{prev}(u_0)}{DC_{prev}(u_0)}
\]

Output values and their meanings are shown in table 5-1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Status</th>
<th>Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Influence</td>
<td>1: Below Average</td>
<td>(PR(u) &lt; PR_{avg})</td>
</tr>
<tr>
<td></td>
<td>2: Constant</td>
<td>(PR(u) = PR_{avg})</td>
</tr>
<tr>
<td></td>
<td>3: Above Average</td>
<td>(PR(u) &gt; PR_{avg})</td>
</tr>
<tr>
<td>Change in Social Influence</td>
<td>1: Decreasing</td>
<td>(PR_r(u) &lt; PR_{r,prev}(u))</td>
</tr>
<tr>
<td></td>
<td>2: Stable</td>
<td>(PR_r(u) = PR_{r,prev}(u))</td>
</tr>
<tr>
<td></td>
<td>3: Increasing</td>
<td>(PR_r(u) &gt; PR_{r,prev}(u))</td>
</tr>
<tr>
<td>Content Relevance</td>
<td>1: Mostly Irrelevant</td>
<td>(CR(u) &lt; CR_{avg})</td>
</tr>
<tr>
<td></td>
<td>2: Neutral</td>
<td>(CR(u) = CR_{avg})</td>
</tr>
<tr>
<td></td>
<td>3: Mostly Relevant</td>
<td>(CR(u) &gt; CR_{avg})</td>
</tr>
<tr>
<td>Change in Content Relevance</td>
<td>1: Decreasing</td>
<td>(CR_s(u) &lt; CR_{s,prev}(u))</td>
</tr>
<tr>
<td></td>
<td>2: Stable</td>
<td>(CR_s(u) = CR_{s,prev}(u))</td>
</tr>
<tr>
<td></td>
<td>3: Increasing</td>
<td>(CR_s(u) &gt; CR_{s,prev}(u))</td>
</tr>
<tr>
<td>Degree Centrality</td>
<td>1: Abnormal</td>
<td>(DC(u) &lt; DC_{avg})</td>
</tr>
<tr>
<td></td>
<td>2: Neutral</td>
<td>(DC(u) = DC_{avg})</td>
</tr>
</tbody>
</table>
In the following, we discuss the detailed input requirements, functionalities and output values of the intermediate nodes.

1. **Influence:**

   Node influence has two inputs, SI and Change in CSI. Table B-6 shows the input values and the corresponding weights. According to the table, both the high SI and high CSI combination and low SI and low CSI combination has the extremely diversified weight configuration. The reason for such setting is for capturing the influential abusive users through this set of nodes. Output values and their meanings are shown in table B-10.

2. **Content:**

   Content has two input nodes, CR and CCR. Content measures value change of in the relevance of user’s posts. Both the combination of high values and low values on the two inputs has extremely high weights respectively. The reason for such setting is also for capturing the highly irrelevant posts from abusive users through this set of nodes. Output values and their meanings are shown in table B-11.
3. Centrality,

Centrality measures if the in-degree centrality and out-degree centrality ratio. It has two input nodes, DC and CDC. The Centrality node examines the balance of the number of posts submitted by the user and the number of replies he received. Output values and their meanings are shown in table B-12.

5.3 Detailed Analysis of Core Model Nodes

In the following, we discuss the detailed input requirements, functionalities and output values of the core model nodes.

1. Intent:

The new Intent node has Content, Authenticity and Activity as inputs. This configuration makes new Intent more powerful. With the add-on of the node Content, Intent now takes actual relevance of the topics in user's posts into account. This aspect enhances the ability of detecting the user’s intention on abusing the forum.

2. Capacity:

The new Capability node has Resource, Access, Influence and Centrality as inputs. The newly joined nodes Influence and Centrality have greatly improved the node Capability on detecting the threat of user, with these two nodes, we now can monitor the user’s ability for abusing the forum to a new level. Especially
with node Influence, the approach of Page Rank ranks the users in the way their popularity spreads.

Besides BN, a plugin tool is included in the solution to help administrators monitor actions in the forum with just a few clicks. There are four major components in the design, the forum, plugin tool, database and server. The plugin is designed for easy installation on top of a forum with self-deployable database.

The plugin has a well design web interface for monitoring all users’ activities in real time. Every time user submits post, the backbone server will run the Bayesian Network and calculate the latest scores for each user. Then it automatically updates the result on the plugin tool. All the results and running status of the plugin will be stored in the database.
Figure 5-4 Components of the Extended Solution System
Chapter 6

Algorithms

This chapter introduces the algorithms implemented in the thesis. In the Bayesian Network described in Chapter 4 and 5, the input nodes execute all of the data processing operations. Such operations require algorithms related to topic mining, sentimental analysis, page rank, mutual information and etc.

There are twelve input nodes, Authentic Behavior, Change in Authentic Behavior, Post Behavior, Change in Post Behavior, Contribution Behavior, Change in Contribution Behavior, Social Influence, Change in Social Influence, Content Relevance, Change in Content Relevance, Degree Centrality and Change in Degree Centrality.

1. Authentic Behavior

The node Authentic Behavior (AB) measures the authenticity of a user’s posts. To compute this value, the node requires all the posts of user as input. The node has two different algorithms for processing the posts. The first one counts the number of cursed words. Swear words are originated from a pre-defined word dictionary [23]. For each post, one algorithm compares the words in the text to this dictionary and calculates the ratio of the number of cursed words against the number of all words for the post. At the end, this algorithm produces the mean ratio of all the posts belonging to the current user. The other algorithm calculates the sentiment value of the text of the posts from a sentiment dictionary [24]. This sentimental dictionary contains a wide variety of lexicons. Each lexicon in the dictionary is mapped to a sentiment value. The sentiment value of each post is the sum of sentiment values of all the
words divided by the number of words in the post. The output of this algorithm is the mean sentiment value of all the posts belonging to the current user. The output of AB is composited by the outcome of both algorithms. The detailed algorithm is described as follows:

```
output = 0
badword_ratio = 0
sentimental_value = 0
load all cursed word from dictionary dA into hash table hA, using the bad words as key
load the sentimental value and lexicons from dictionary dB into hash table hB, using the lexicons as key
load all the texts of posts into array text_array
number_of_post = length of
for each text in text_array
    eliminate all punctuations
delete all stop words
    badword_local = 0
    sentimental_value_local = 0
    split the post into individual words and store them into array words
    for each word in words
        if the word appears in the hash table of bad word
            badword_local++
        if the word appears in the hash table of sentimental dictionary
            sentimental_value_local += the sentimental value of the word
    badword_local /= number_of_word
    sentimental_value_local /= number_of_word
    badword_ratio += badword_local
    sentimental_value += sentimental_value_local
badword_ratio /= number_of_posts
sentimental_value /= number_of_posts
output = 0.5 * badword_ratio + 0.5*sentimental_value
Compares the value to the average of the forum
if the final value is higher than the average
    return 3
else if equal
    return 2
else return 1
```

The time complexity for this algorithm is O(n^2). Since for each post, both algorithms examine each word in the post against dA and dB. Because dA and dB are hash
tables, the examine process takes O(1). To compare the entire post, the time complexity become is O(n). The time complexity for all the posts is O(n^2). For space complexity, the space required are four arrays, two hash tables and other variables, therefore the space complexity is O(n).

2. Change in Authentic Behavior,

Change in Authentic Behavior calculates the difference between the Authentic Behavior values of two different time stamps. In this thesis, the choice of the two time stamp are always the current time stamp and the time stamp of half the time elapsed from the beginning until the current time stamp.

```
Initialize variable a1
Initialize variable a2
a1 = Authentic Behavior(current timestamp)
a2 = Authentic Behavior((current timestamp - beginning timestamp) / 2)
if a1 > a2
    return 3,
else if a1 == a2
    return 2
else return 1
```

The time complexity for this algorithm is O(n^2), since it has executed two AB operations. Space complexity is O(n).

3. Post Behavior

Post Behavior measures the volume of posts submitted by a user against the average value of all other users. This value indicates the activeness of the user. If the Post Behavior value of a user is higher than the mean of the forum in a period of time, then the user is likely to be more active than others. The node calculates the value by
measuring the total number of posts submitted by the user of the current time and compares it with the mean value of the forum. The detailed algorithm is shown as follows:

```python
number_of_user_post = 0
number_of_average_post = 0
number_of_user_post = query count(*) of posts belongs to user
number_of_average_post = query count(*) of total number of post / query count(*) of total number of user
if number_of_user_post > number_of_average_post value is higher
    return 3
else if number_of_user_post == number_of_average_post
    return 2
else return 1
```

The time complexity for this algorithm is $O(1)$, since all the operations executed are just a limited number of queries to the database. Space complexity is $O(1)$.

4. Change in Post Behavior

Change in Post Behavior calculates the difference between the Post Behavior values of two different time stamps. In this thesis, the choice of the two time stamp are always the current time stamp and the time stamp of half the time elapsed from the beginning until the current time stamp.

```python
Initialize variable a1
Initialize variable a2
a1 = Post Behavior(current timestamp)
a2 = Post Behavior((current timestamp - beginning timestamp) / 2)
if a1 > a2
    return 3,
else if a1 == a2
    return 2
else return 1
```

The time efficiency for this algorithm is $O(1)$ and space efficiency is $O(1)$. 
5. Contribution Behavior

Contribution Behavior measures overall contribution of user. Such value is measured by calculating Authentic Behavior values by each thread the user has contributed to. The raw material for the node is also the posts and a time stamp. To calculate the value, the values of Authentic Behavior are measured by thread. For example, if user posts seven posts on thread A1 and eight posts on thread B1, then the Contribution Behavior is calculated by summing the total Authentic Behavior value and divide it by the number of threads, and in this case is two. Then, similar procedure runs on all posts and forms an outcome as the detailed algorithm shown below:

```
output = 0
load all posts for user to array posts
categorize them by threadids
for all threads
    contribution_behavior_local = 0
    number_of_post_by_thread = count(threadid);
    for each post in the thread
        contribution_behavior_local += Authentic Behavior(post)
    contribution_behavior_local /= number_of_post_by_thread
    output += contribution_behavior_local
output /= number_of_thread

average = 0
load all posts from database
categorize them by threadids
for all threads
    contribution_behavior_local = 0
    number_of_post_by_thread = count(threadid);
    for each post in the thread
        contribution_behavior_local += Authentic Behavior(post)
    contribution_behavior_local /= number_of_post_by_thread
    average += contribution_behavior_local
average /= number_of_thread

if output > average
    return 3
else if output == average
    return 2
else return 1
```
The time complexity for this algorithm is $O(n^2)$, since it has executed $n$ Authentic Behavior operations and space complexity is $O(n)$.

6. Change in Contribution Behavior

Change in Contribution Behavior calculates the difference between the Post Behavior values of two different time stamps. In this thesis, the choice of the two time stamp are always the current time stamp and the time stamp of half the time elapsed from the beginning until the current time stamp.

```python
Initialize variable a1
Initialize variable a2
a1 = Contribution Behavior(current timestamp)
a2 = Contribution Behavior((current timestamp - beginning timestamp) / 2)
if a1 > a2
    return 3,
else if a1 == a2
    return 2
else return 1
```

The time complexity for this algorithm is $O(n^2)$ and space complexity is $O(n)$.

7. Social Influence

Social Influence measures the quantitative influence the user has upon the entire forum. The algorithm implemented in this node is Page Rank. Consider the forum is a graph. The users are nodes. Posts and replies are edges. For a given time stamp the algorithm Page Rank measures the number of replies each user has and applies these values to Eq. 3-5. In the actual algorithm for computing this value, the node retrieves the list of repliers. After acquiring all the replier information, it acquires the number
repliers for each unique replier. The detailed algorithm is shown in the following pseudo code snippet:

```java
output = 0
load all the user post into array posts
initialize hash table repliers, use userid as key and link_count as value
for each post in posts
    load the repliers into array rp
    if rp is not in repliers
        put rp in repliers
        and link_count of rp as value

base = 1 / size_of_repliers
for each element in repliers
    output += base / link_count

average_rp = load average rp from database

if output > average_rp
    return 3
else if output == average_rp
    return 2
else
    return 1
```

The time complexity for this algorithm is O(n^2), since all the repliers is examined. There are n posts for the user. The space complexity is O(n) since there is a hash table to store.

8. Change in Social Influence

Change in Social Influence calculates the difference between the Post Behavior values of two different time stamps. In this thesis, the choice of the two time stamp are always the current time stamp and the time stamp of half the time elapsed from the beginning until the current time stamp.
Initialize variable a1
Initialize variable a2
a1 = Social Influence(current timestamp)
a2 = Social Influence((current timestamp - beginning timestamp) / 2)
if a1 > a2
  return 3,
else if a1 == a2
  return 2
else return 1

The time complexity for this algorithm is $O(n^2)$, because two Social Influence operations are executed and space complexity is $O(n)$.

9. Degree Centrality
Degree Centrality measures the ratio of the number of posts submitted by the user and the number of replies he received. This node monitors the balance of the user’s post in and out ratio. To calculate this value, the node gathers the number of posts and the number of replies of the user. Divide the first value to the second value.

Please refer to the detailed algorithm in the following pseudo code snippet:

number_of_posts = query count(*) of posts for the user
number_of_replies = query count(*) applies for the user
centrality = number_of_replies / number_of_posts
average_centrality = query average centrality from database

if centrality > average_centrality
  return 3
else if centrality == average_centrality
  return 2
else
  return 1
The time complexity for this algorithm is $O(1)$, since only a few queries are executed. The space complexity is $O(1)$.

10. Change in Degree Centrality

Change in Degree Centrality calculates the difference between the Post Behavior values of two different time stamps. In this thesis, the choice of the two time stamp are always the current time stamp and the time stamp of half the time elapsed from the beginning until the current time stamp.

```python
Initialize variable a1
Initialize variable a2
a1 = Degree Centrality(current timestamp)
a2 = Degree Centrality((current timestamp - beginning timestamp) / 2)
if a1 > a2
    return 3,
else if a1 == a2
    return 2
else return 1
```

The time complexity for this algorithm is $O(1)$ and space complexity is $O(1)$.

11. Content Relevance

Content Relevance measures the relevance ratio of the user’s posts against the threads he has contributed to. To calculate this value, the wordnet dictionary [25] is used to provide synonym references. The wordnet dictionary contains a wide variety of words. The words in the dictionary are categorized into synonym sets (aka synset). At the beginning, CR will cut posts into tokens. Then punctuations and stop words are removed from the token list. After this process, the remaining tokens will be sent to a word processor to remove all the postfixes. For instance, postfixes like –al, -ous
are removed. The processed tokens contain only the core lexical. Later, these tokens are queried in the wordnet dictionary to form a list of synsets. Each synset represents a group of synonyms. If the token does not appear in wordnet dictionary, it will be added to a list of unrecognized words. After processing the post, the very same procedure is run on the thread the post associated with. Now two lists of synsets and unrecognized words are produced for the post and the thread. The list is used to calculate mutual information between the post and the thread. Mutual information described in Chapter 3 measures the similarity presented by two different messages.

The detailed algorithm is described as follows.

```plaintext
output = 0  # Initialize the hash table hA for new words, using the words as keys.
load wordnet into hash table hB, using the lemma as key and synsetids as value.
get all the posts for the user and load them into array posts.
for each post in posts:
    remove all redundancies.
    split the post into words and store in array w.
    initialize array synset_post.
    for each word:
        if the wordnet contains the word
            add the synsetid to synset aday.
    get all the posts that have the same threaded as the current post.
    an load into array thread.
    get the synset array of thread.
    output_local = calculate the mutual information of the post and the thread using the synset array of both.
    output += output_local.
    output /= number_of_threads.

cr_average = query from database.
if output > cr_average:
    return 3.
else if output == cr_average:
    return 2.
else:
    return 1.
```
The time complexity for this algorithm is \(O(n^2)\). Since for each post and thread, all the words are examined. There are \(n\) posts and threads. The space complexity is \(O(n)\), because there are a number of arrays and hash tables to store.

12. Change in Content Relevance

Change in Content Relevance calculates the difference between the Post Behavior values of two different time stamps. In this thesis, the choice of the two time stamp are always the current time stamp and the time stamp of half the time elapsed from the beginning until the current time stamp.

```plaintext
Initialize variable a1
Initialize variable a2
a1 = Content Relevance(current timestamp)
a2 = Content Relevance((current timestamp - beginning timestamp) / 2)
if a1 > a2
   return 3,
else if a1 == a2
   return 2
else return 1
```

The time complexity for this algorithm is \(O(n^2)\), since two Content Relevance operations are executed. The space complexity is \(O(n)\).
Chapter 7

Design and Implementation

The system includes three major components, database which holds data of forum, plugin tool for displaying results and backbone system for running calculations. This chapter introduces design and implementation details for plugin tool and backbone system.

7.1 Backbone System

The backbone system is implemented in Java and Matlab. This part of the system is in charge of gathering inputs and coordinates the operations for calculating the final result. The input gathering nodes are implemented in Java. Java has well designed interface towards database and Matlab. Bayesian Network is built using a tool package for Matlab named FullBNT, for the face that Matlab is faster for manipulating matrixes.

The Java application contains eight building blocks, IO package, Key management package, Basic data structures package, Exceptions package, Data Management package, Nodes package, Word Processor package and Procedure package.

IO package contains five classes, IOFactory class, IO class, MatlabIO class, DatabaseIO class and FileIO class. The IO component of the entire system uses the Factory Method design pattern to allow easy access to the IO technologies. Factory Method pattern allows accessing entirely different sources of IO input by just using different parameters but with the same function names. Therefore, once implemented, the other classes that require certain IO sources can access them without knowing the underlying details of the IOFactory class. When modification is needed for better performance, only the individual IO classes need to be changed. The DatabaseIO class requires the Key management package, which contains the keys and
connection parameters of the databases, whenever a call for making a database connection is initiated, the Key management component will also instantiated. Through this Key management component, the IO classes can easily retrieve the keys for authenticating with databases. The class diagram of IO package is shown in Figure 7-1 and the class diagram of Key management package is shown in Figure 7-2.
The basic data structure package contains nine classes, Post class, Posts class, Replies class, Thread class, Lexicon class, Unrecognized class, SynSet class, User class and PostUtility class.

Post class is an abstract concept of a data structure for holding a post and it has encapsulates all the essential fields for posts stored in the post table in the database. These fields include postid, userid, text, parentid, threadid, timeline. Each post in the forum has a unique postid, and each post is associated to a userid. The parentid of the post refers to the post that this post is replying to and the threadid is the thread the current post associated with. The timeline field indicates when the post is submitted to the forum.

Posts class encapsulates all the posts for a userid. This class contains a userid, a hash table for storing all the posts, and a size value storing the number of all the posts in the hash table. The hash table in the Posts class uses the postids for key and Post instances for values.

Replies and Thread class is similar to the Posts class, the difference of the Replies class is this class stores all the posts that has been replied to a postid. Therefore instead of storing a userid, hash table and size the Replies class has a postid, hash table and size. The Thread class contains the threadid, the hash table storing all the Posts with the same threadid and the size.
Lexicon class stores the sentimental lexicon with their corresponding positive and negative sentimental values. The Lexicon contains a hash table using word as key and a double array as value to store the positive and negative value pair, and the size of the Lexicon.

SynSet stores the synset mapping of the wordnet. This data structure contains a hash table using word as key and an array to store all the synsetids and the size of the SynSet.

Unrecognized class is used to keep the record of the words that are not in the SynSet, it contains exactly same elements as SynSet.

User class stores all the information of this user, it contains field such as userid, username, screen name, join date, current privilege level.

Finally the PostUtil class, this class contains the methods than can transform any of the data structures above into array or backwards, and most importantly, this class contains the methods that can shrink the classes like Posts, Replies and Threads by an exact number of posts, by ratio or to a date.

The design of these classes contains mostly hash tables, the hash tables are known for O(1) access and O(log n) comparison. Using hash tables than just arrays can greatly optimize the performance when calculating Content Relevance and Page Rank. The classes in this package defines clean and neat access of the input nodes to access any information required. Figure 7-3
shows the class diagram of this package.

Figure 7-3 Basic Data Structure Package Diagram

Package exception provides customized exception classes that cover error handling to make sure the execution goes smoothly and automatically. There are seven exceptions in this package, NoRepliesException class, UndefinedLexiconException class, UndefinedSynSetException class, InvalidShrinkRatioException class, InvalidShrinkDateException class, InvalidShrinkSizeException class, InvalidTypeException class.

NoRepliesException class is thrown when no replies for the current user is found.

UndefinedLexiconException class is thrown when the word trying to look for is not in Lexicon.

UndefinedSynSetException class is thrown when the word is not identical to any of the lemma in the SynSet. InvalidTypeException class is thrown when the method for processing Posts, Replies
or Thread is taken a wrong type of parameter. The InvalidShrinkRatioException class is thrown when the shrink ratio for a Posts, Thread or Replies is unable to shrink the size by at least 1. The InvalidShrinkDateException class is thrown when the date designated to shrink is earlier than the earliest date in the Posts, Replies or Thread. The InvalidShrinkNumberException class is thrown when the number of post to shrink is greater than the size of Posts, Replies or Thread.

![Exception Package Class Diagram](image)

Figure 7-4 Exception Package Class Diagram

By defining these exceptions, the programs became more flexible when encountering abnormal conditions. Figure 7-4 shows the class diagram.

Data management package contains only one class, the DataFactory class. This class handles all the data fetching request. This class contains method for fetching the Posts, Thread, Replies, Lexicon, SynSet and User data structures from the database. All the input nodes only need to call functions in this class to get the data required. Figure 7-5 shows the class diagram.
The package containing the most classes is the nodes package. There are sixteen classes, AuthenticBehavior class, ChangeInAuthenticBehavior class, ContributionBehavior class, ChangeInContributionBehavior class, PostBehavior class, ChangeInPostBehavior class, PageRank class, ChangeInPageRank class, ContentRelevance class, ChangeInContentRelevance class, DegreeCentrality class, ChangeInDegreeCentrality class, TargetPost class, TargetUser class, PreviousSanction class and ForumAttention class. Each of the class operates as described in the Chapter 4, 5 and 6. The class diagram is shown in Figure 7-6.

PageRank node is optimized for better performance. A set of page rank values is calculated in the beginning of the test for all the users and first 10 days of all the posts and stored as file. When doing PageRank calculations, only the elements with changes is recalculated and put into the formula.

Word Processor package contains four classes, PunctuationRemover class, StopWordRemover class, WordStemmer class, VectorBuilder class and WordCleaner class. The
PunctuationRemover breaks the word into characters and remove the punctuation character and reassemble the word. The StopWordRemover removes the stop words and the

WordStemmer checks all the postfix in record and remove the postfix of the word. The

WordCleaner class has the function the bundle all the previous three class and clean the word
once and for all. The VectorBuilder class has a hash table that stores the cleaned word as key and the frequency as value. Figure 7-7 shows the class diagram.

Procedure package contains class for the sequence of executing the program. There is an entry point class Exec, the Initialization class, the GetInputs class and GetResult class. The Exec will call Initialization which initializes the Lexicon, SynSet and Posts. Then the GetInput class is called to get all the input value into an array. After that, the GetResult class uses this array to fetch the threat value from Matlab. Figure 7-8 shows the class diagram.

![Procedure Page Class Diagram](image)

Figure 7-8 Procedure Page Class Diagram

After the system gathered the input, the data is send to Matlab via Java Remote Method Call. This is a way to let Java control applications running on another process and exchange data with this process. The library matlabcontrol is created by MathWorks to allow Java sending command and retrieve data from Matlab through their MatlabProxy interface. The actual code for building the Bayesian Network is provided by the library FullBNT.
7.2 Plugin Tool

Another component of this thesis is the plugin tool. The plugin tool is implemented using the latest UI technology HTML5, SVG, CSS3, Ajax and JQuery. And to accommodate the backbone system in Java, the interface between UI and backbone system is Node.js Server.

The UI contains two main screens. The first screen is the login portal. Once the administrator logged onto the tool, the screenshot of the second main screen is shown as in figure 7-9. There are two parts of the screen, a user list is shown on left and the charts on threat value changes regarding to different Intent, Opportunity and Capability value. And the user of this tool can also change the view to a number of other metrics by clicking the tab below. Other views include the F-rate, values in other nodes through time. Other than the actual user results shown,
there is also a cumulated trend in the user list that will show the average value changes in the entire forum. The basic layout of the UI uses HTML5. HTML5 is the base engine for rendering the charts in canvas, and can be read from any modern browsers. The CSS3 describes the positions, colors, shadows and animations. JQuery is a library for Javascript, which allows objected orient programming in Javascript. Combining JQuery, Ajax and Node.js, the plugin is able to retrieve real time results from the database and dynamically updates the charts without refreshing the whole page.

The logic flow of the plugin is designed as follows. First the user click a userid listed on the top part of the UI. Then JQuery captures the click event. JQuery then dispatches an HTTP request to Node.js. Node.js monitors port 8888 to 7776 and capture the request. The node.js will run the logics encapsulated inside and retrieve the information requested. To return, the receivers encapsulate the data in a JSON file and send back to the web front. In the last step of the HTTP request, the status 200 indicates that data is successfully returned. Upon checking the status code, JQuery decodes the JSON file and extract several arrays of data. It translates this data into drawing points. Finally the new chart is updated on the right part of the screen.
Chapter 8
Test and Result

The test data used for our evaluation is the Zelda Vbull Gaming forum data image. Zelda Vbull is a very famous gaming forum with more than 10,000 users and 600-million posts in total. The site also contains a large numbers of abusive posts and spammers. Abusive and spammer posts are stored in a dedicated table of the Zelda database, referred to as infraction table.

The test case contains 50 users with more than 8,000 posts per user. Based on observation, we noticed that these users are not those being frequently moderated. Therefore, to execute the test more accurately, artificial factors are added to the test case where we randomly inject malicious posts to each user. The injected posts are randomly selected from the infraction table in the forum database.

The detailed testing strategy includes two sets of tests. These tests are designed to compare the performance difference between the original Bayesian Network and the new one.

In both tests, for each user, 30% moderated posts out of all posts per user are randomly injected. For instance, if a user has 10,000 posts, then an additional 3,000 moderated posts are randomly inserted to user’s post bank. In the test, three metrics is selected for tests, Sensitivity, Specificity and the F-rate.

The equation for calculating F-rate is show below.

\[
\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}
\]

\[
\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}
\]

\[
F\text{-rate} = 2 \times \frac{\text{specificity} \times \text{sensitivity}}{\text{specificity} + \text{sensitivity}}
\]
Intuitively, for any test, there is usually a trade-off between the measures. For example: in an airport security setting in which one is testing for potential threats to safety, scanners may be set to trigger on low-risk items like belt buckles and keys (low specificity), in order to reduce the risk of missing objects that do pose a threat to the aircraft and those aboard (high sensitivity). This trade-off can be represented graphically as a receiver operating characteristic curve.[52]

![Figure 8-1 Test Result of Trico (Core Solution)](image)

In Figure 8-1, the blue curve represents the F-rate of the original Bayesian Network, and the red curve shows the value given from a human moderator. Given the results shown in Figure 8-1, Bayesian network yields much higher number of true positives and true negatives than human moderator, at which point, a conclusion can be made that the Bayesian network performs much better than human moderator in terms of both accuracy and overall coverage of posts.

Before the actual test, training is required to properly tune the TriCO conditional probability tables and the system parameters. The first training process is to run the test on only posts in the infraction table. The purpose of this test is to set the threshold of the threat value to a proper bar, so that the threshold is high enough for yielding high rate of true positive results.
After each round of training, if the threshold is not in the desired range, then the weights in the Conditional Probability Table will need to be changed. In the test for this thesis, totally eight rounds of such test is conducted. In the end, the threshold for dividing good and bad post is 0.79813725. The second training process runs for each user tested. This training process takes the first fifty posts out of the all the posts belonging to the current user. In exact terms, if a user has 8000 posts in total, the from post number one to post number fifty, these posts are used for training. And from the post number fifty-one to eight-thousand, these posts are used for generating the threat value.

The test took 36 hours 53 minutes and 23 seconds to finish on a quad-core machine with 3.2GHz clock rate. Memory used is 5,672Mb max and 4,311 Mb on average.

Figure 8-2 to 8-4 shows the result comparison for just one user. This user has a total number of 11,315 posts that are originated by him. 3,394 malicious posts are injected to him, therefore, now he has 14,709 posts. As can be concluded from the figures, the specificity values of the new Bayesian Network are noticeably higher than the original network, while the sensitivity values are higher than for a lower margin. In other words, the total number true positives are similar but the number true negatives are largely different. This kind of combination results in a mild difference of f-rate values. Table 8-1, shows the statistical difference between the two Bayesian Networks. For this one user, we can conclude that the new Bayesian Network performs better than the original network.

Now for the average result of all the 50 users tested, the difference of the three metrics shows some even better results than single-user results shown in the previous three figures.
Figure 8-2 Test Result Comparison for One User (Sensitivity)
Figure 8-3 Test Results Comparison for One User (Specificity)

Figure 8-4 Test Results for One User (F-rate)
Table 8-1 Quantified Test Results Comparison for One User

<table>
<thead>
<tr>
<th>New BN over Original BN</th>
<th>Max Difference</th>
<th>Min Difference</th>
<th>Mean Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.0712384</td>
<td>-0.0312391</td>
<td>0.02318231</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.1712931</td>
<td>-0.02930121</td>
<td>0.151012903</td>
</tr>
<tr>
<td>F-rate</td>
<td>0.19231890</td>
<td>-0.12301030</td>
<td>0.123212841</td>
</tr>
</tbody>
</table>

By comparing the results reported in Table 8-2, we conclude not only that the new Bayesian Network performs better than the original Bayesian Network on the all three metrics for the mean value across all users, but also that the difference in value is higher and more stable. The F-rate values from the new Bayesian Network are all higher than the original Bayesian Network.

![Sensitivity Mean](image)

Figure 8-5 Test Results Comparison for Average of Fifty Users (Sensitivity)
Figure 8-6 Test Results Comparison for Average of Fifty Users (Specificity)

Figure 8-7 Test Results Comparison for Average of Fifty Users (F-rate)
Table 8-2 Quantified Test Results Comparison for Average for Fifty Users

<table>
<thead>
<tr>
<th>New BN over Original BN</th>
<th>Max Difference</th>
<th>Min Difference</th>
<th>Mean Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.12813871</td>
<td>-0.000123123</td>
<td>0.1034231</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.561123341</td>
<td>0.12930121</td>
<td>0.3410446</td>
</tr>
<tr>
<td>F-rate</td>
<td>0.34981274</td>
<td>0.192031201</td>
<td>0.253212841</td>
</tr>
</tbody>
</table>

Performance wise, the new Bayesian Network outperforms the original one. System performance is also measured against the two networks. The response time measured only includes the calculation time with the data in machine memory, and the data access time from database is not taken into account. The reason is the time for accessing the database is different from the time of calculation by at least two orders of magnitudes. Figure 8-8 shows the mean response time of per user per round for all the 50 users tested. As can be concluded from the figure, the new Bayesian Networks takes on average 1000ms more than the original one. This is understandable since the new network includes more nodes and processes more information. For example, some nodes like the Content Relevance and Page Rank can take considerably more time to compute. Yet, the response time is still in an acceptable range, since the difference is merely one second.

To evaluate the tests, first of all the tests are set to have 30% of the malicious posts injected. Second, the threshold for detecting the malicious post is set by continuously train the BN using malicious posts only. Given the current setting the BN and test design, new BN is able to provide more accurate predictions than before. Overall, the new BN has a very good performance in detecting both normal posts and malicious posts. In reality, the less malicious
posts a user has out of total number of posts the user submits, the more difficult to predict the user’s harmfulness. Then, the forms and contexts of the malicious posts can be different than the ones stored in the infraction table.

![System Performance in Milliseconds](image)

**Figure 8-8 Average Performance Comparison**
Chapter 9

Future Works

For future work, there are varieties of ways to improve the approach introduced in the thesis. First, to further explore the social factor of the network, the Social Influence node can be expanded from an input node to a core model node. Instead of just measuring the number of replies as ranking factor, the node can count the number of views and the number of friends. The Content Relevance can also be used as additional weight for the ranking. Machine learning is a useful tool that can be added to the system, letting the system to adapt to the changes in the social community to change the weights in the CPTs to deliver better results specially tuned for the community. Better sentimental analysis, content relevance algorithms other than just mutual information can be implemented to improve the accuracy on input nodes, e.g., Authentic Behavior and Content Relevance.

Other social analysis method can be used as inputs such as closeness centrality and betweenness centrality.

The system can be modified to accommodate other SNS such as Facebook, Twitter, Renren by implementing interface to the APIs provided from third-party.
Chapter 10

Conclusion

This thesis has created the first try of using both Bayesian Belief Network and Page Rank method to find policy violators in an online social community. After a sufficient amount of test, the result has shown much better accuracy and performance than human moderators. The thesis has also created plugin tool for Bulletin Board System technology.

This thesis is a start point on my adventure on social related data mining. This is a very interesting and thriving field. I shall continue on working other useful tools to change the lives of others.
Reference


redesign/?awesm=tnw.to_1E09r&utm_campaign=social%20media&utm_medium=copy-paste-
link&utm_source=referral&utm_content=Now%20with%20170%2320million%20users,%20Google%20%202. [Accessed 16 April 2012].


[Accessed 16 April 2012].


pp. 211 - 220.


[33] M. Marcelo, A. Jussara and A. Virgilio, "Identifying user behavior in online social networks," in SocialNets '08: Proceedings of the 1st Workshop on Social Network


Table B-1 CPT for Character

<table>
<thead>
<tr>
<th>Variable</th>
<th>CCR</th>
<th>CR</th>
<th>CONTAINED</th>
<th>SPONTANEOUS</th>
<th>PERSISTENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1 1</td>
<td>Decr</td>
<td>an</td>
<td>Below</td>
<td>Ave</td>
<td>3 2 1</td>
</tr>
<tr>
<td></td>
<td>aisi</td>
<td>d  le</td>
<td>ow</td>
<td>g e</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 1</td>
<td>Stable</td>
<td>a</td>
<td>Below</td>
<td>Ave</td>
<td>3 2 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dle</td>
<td>ow</td>
<td>g e</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 1</td>
<td>Incr</td>
<td>a</td>
<td>Below</td>
<td>Ave</td>
<td>2 1 3</td>
</tr>
<tr>
<td></td>
<td>aisi</td>
<td>dle</td>
<td>ow</td>
<td>g e</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 2</td>
<td>Decr</td>
<td>a</td>
<td>Ave</td>
<td>rag e</td>
<td>3 1 2</td>
</tr>
<tr>
<td></td>
<td>aisi</td>
<td>dle</td>
<td></td>
<td>e</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 2</td>
<td>Stable</td>
<td>a</td>
<td>Ave</td>
<td>rag e</td>
<td>2 1 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dle</td>
<td></td>
<td>e</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 2</td>
<td>Incr</td>
<td>a</td>
<td>Ave</td>
<td>rag e</td>
<td>1 2 3</td>
</tr>
<tr>
<td></td>
<td>aisi</td>
<td>dle</td>
<td></td>
<td>e</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Costs (19)</td>
<td>CPT for Cost</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>--------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Major Cost</td>
<td>Some Cost</td>
<td>No Cost</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 2 3</td>
<td>How much more likely is MAJOR COST than SOME COST?</td>
<td>2</td>
<td>How much more likely is SOME COST than NO COST?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.6</td>
<td>0.0</td>
<td>0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 2 1</td>
<td>How much more likely is NO COST than SOME COST?</td>
<td>1</td>
<td>How much more likely is SOME COST than MAJOR COST?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 1 2</td>
<td>How much more likely is SOME COST than NO COST?</td>
<td>1</td>
<td>How much more likely is NO COST than MAJOR COST?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 2 3</td>
<td>How much more likely is MAJOR COST than SOME COST?</td>
<td>4</td>
<td>How much more likely is SOME COST than NO COST?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 2 3</td>
<td>How much more likely is MAJOR COST than SOME COST?</td>
<td>4</td>
<td>How much more likely is SOME COST than NO COST?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 2 1</td>
<td>How much more likely is NO COST than SOME</td>
<td>1</td>
<td>How much more likely is SOME COST than MAJOR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table B-03 CPT for Benefit

<table>
<thead>
<tr>
<th>Variable</th>
<th>Character</th>
<th>Target Posts</th>
<th>Target Users</th>
<th>MAJOR BENEFIT</th>
<th>SOME BENEFIT</th>
<th>NO BENEFIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>COST?</td>
<td>F</td>
<td>COST?</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>COST?</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BENE</td>
<td>FITS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BENE</td>
<td>FITS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BENE</td>
<td>FITS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BENE</td>
<td>FITS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>1</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Character</th>
<th>Target Posts</th>
<th>Target Users</th>
<th>MAJOR BENEFIT</th>
<th>SOME BENEFIT</th>
<th>NO BENEFIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTAINED</td>
<td>Attactive Posts</td>
<td>and</td>
<td>Attactive Users</td>
<td>and</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>SPONTANEOUS</td>
<td>Attactive Posts</td>
<td>and</td>
<td>Attactive Users</td>
<td>and</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>PERSISTENT</td>
<td>Attactive Posts</td>
<td>and</td>
<td>Attactive Users</td>
<td>and</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>Contained</td>
<td>NO Posts</td>
<td>Attractive Users</td>
<td>1</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>-----------</td>
<td>---------</td>
<td>----------------</td>
<td>---</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>Spontaneous</td>
<td>NO Posts</td>
<td>Attractive Users</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>Persistent</td>
<td>NO Posts</td>
<td>Attractive Users</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>Contained</td>
<td>NO Users</td>
<td>Attractive Posts</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>Spontaneous</td>
<td>NO Users</td>
<td>Attractive Posts</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>Persistent</td>
<td>NO Users</td>
<td>Attractive Posts</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>Contained</td>
<td>NO Users</td>
<td>NO Users</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>Spontaneous</td>
<td>NO Users</td>
<td>NO Users</td>
<td>3</td>
</tr>
<tr>
<td>ntaneous</td>
<td>Persistent</td>
<td>NO Posts</td>
<td>NO Users</td>
<td>Benefit than SOME Benefit?</td>
<td>Benefit than MAJOR Benefit?</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td></td>
</tr>
<tr>
<td>322</td>
<td>312</td>
<td>5</td>
<td></td>
<td>4970</td>
<td>0.900.810</td>
<td></td>
</tr>
</tbody>
</table>

Table B-4 CPT for Authenticity

<table>
<thead>
<tr>
<th>Variable</th>
<th>CAR</th>
<th>BR</th>
<th>INAUTHENTIC</th>
<th>AUTHENTIC</th>
<th>N/A</th>
<th>ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Decr</td>
<td>Beloow Average</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>How much more likely is INAUTHENTIC than AUTHENTIC?</td>
<td>10</td>
<td>How much more likely is AUTHENTIC than N/A?</td>
<td>INF</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>Stable</td>
<td>Beloow Average</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>How much more likely is INAUTHENTIC than AUTHENTIC?</td>
<td>20</td>
<td>How much more likely is AUTHENTIC than N/A?</td>
<td>INF</td>
<td>0</td>
</tr>
<tr>
<td>31</td>
<td>Increasing</td>
<td>Beloow Average</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>How much more likely is INAUTHENTIC than AUTHENTIC?</td>
<td>10</td>
<td>How much more likely is AUTHENTIC than N/A?</td>
<td>INF</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>Decr</td>
<td>Average</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>How much more likely is</td>
<td>2</td>
<td>How much more likely</td>
<td>INF</td>
<td>0</td>
</tr>
<tr>
<td>Activity</td>
<td>CPT for Activity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACTIVITY</td>
<td>活性 (RPB, 8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table B-5
<table>
<thead>
<tr>
<th>Variable</th>
<th>CPB</th>
<th>PR</th>
<th>INACTIVE</th>
<th>ACTIVE</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID 2 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 1 Decreasing</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Below Average</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>How much more likely is INACTIVE than ACTIVE?</td>
<td>1</td>
</tr>
<tr>
<td>2 1 Stable</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>How much more likely is INACTIVE than ACTIVE?</td>
<td>2</td>
</tr>
<tr>
<td>Below Average</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>How much more likely is INACTIVE than ACTIVE?</td>
<td>1</td>
</tr>
<tr>
<td>3 1 Increasing</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>How much more likely is INACTIVE than ACTIVE?</td>
<td>3</td>
</tr>
<tr>
<td>Below Average</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>How much more likely is INACTIVE than ACTIVE?</td>
<td>1</td>
</tr>
<tr>
<td>1 2 Decreasing</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>How much more likely is ACTIVE than INACTIVE?</td>
<td>1</td>
</tr>
<tr>
<td>Average</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>How much more likely is ACTIVE than INACTIVE?</td>
<td>2</td>
</tr>
<tr>
<td>2 2 Stable</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>How much more likely is ACTIVE than INACTIVE?</td>
<td>2</td>
</tr>
<tr>
<td>Average</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>How much more likely is ACTIVE than INACTIVE?</td>
<td>5</td>
</tr>
<tr>
<td>3 2 Increasing</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>How much more likely is ACTIVE than INACTIVE?</td>
<td>3</td>
</tr>
<tr>
<td>Above Average</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>How much more likely is ACTIVE than INACTIVE?</td>
<td>3</td>
</tr>
<tr>
<td>1 3 Decreasing</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>How much more likely is ACTIVE than INACTIVE?</td>
<td>1</td>
</tr>
<tr>
<td>Above Average</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>How much more likely is ACTIVE than INACTIVE?</td>
<td>1</td>
</tr>
<tr>
<td>2 3 Stable</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>How much more likely is ACTIVE than INACTIVE?</td>
<td>1</td>
</tr>
<tr>
<td>Above Average</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>How much more likely is ACTIVE than INACTIVE?</td>
<td>2</td>
</tr>
<tr>
<td>3 3 Increasing</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>How much more likely is ACTIVE than INACTIVE?</td>
<td>2</td>
</tr>
<tr>
<td>Variable</td>
<td>Activity</td>
<td>Authenticity</td>
<td>Content</td>
<td>INTENT</td>
<td>NO INTENT</td>
</tr>
<tr>
<td>----------</td>
<td>----------</td>
<td>--------------</td>
<td>---------</td>
<td>--------</td>
<td>-----------</td>
</tr>
<tr>
<td>1 1 1</td>
<td>Inactive</td>
<td>Inactive</td>
<td>Irrelevant</td>
<td>1 2 3</td>
<td>How much more likely is INTENT than NO INTENT?</td>
</tr>
<tr>
<td>2 1 1</td>
<td>Active</td>
<td>Inactive</td>
<td>Irrelevant</td>
<td>1 2 3</td>
<td>How much more likely is INTENT than NO INTENT?</td>
</tr>
<tr>
<td>1 2 1</td>
<td>Inactive</td>
<td>Inactive</td>
<td>Irrelevant</td>
<td>1 2 3</td>
<td>How much more likely is INTENT than NO INTENT?</td>
</tr>
<tr>
<td>2 2 1</td>
<td>Active</td>
<td>Authentic</td>
<td>Irrelevant</td>
<td>1 2 3</td>
<td>How much more likely is INTENT than NO INTENT?</td>
</tr>
<tr>
<td>1 1 2</td>
<td>Inactive</td>
<td>Inactive</td>
<td>Relevant</td>
<td>1 2 3</td>
<td>How much more likely is INTENT than NO INTENT?</td>
</tr>
<tr>
<td>2 1 2</td>
<td>Active</td>
<td>Inactive</td>
<td>Relevant</td>
<td>1 2 3</td>
<td>How much more likely is INTENT than NO INTENT?</td>
</tr>
</tbody>
</table>

Table B-6 CPT for Intent
<table>
<thead>
<tr>
<th>ID</th>
<th>Opportunity</th>
<th>Activity and Authentic and Relevant</th>
<th>2</th>
<th>1</th>
<th>3</th>
<th>How much more likely is NO INTENT than INTENT?</th>
<th>1</th>
<th>0</th>
<th>How much more likely is INTENT than NA?</th>
<th>INF</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>I and Active and Authentic and Relevant</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>How much more likely is NO INTENT than INTENT?</td>
<td>1</td>
<td>0</td>
<td>How much more likely is INTENT than NA?</td>
<td>INF</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table B-7 CPT for Opportunity

<table>
<thead>
<tr>
<th>Variable</th>
<th>BENEFITS</th>
<th>COSTS</th>
<th>OPPORTUNITY</th>
<th>NO OPPORTUNITY</th>
</tr>
</thead>
</table>
| ID 1 1  | MAJORBenefit | MAJORCost | 1 2 | How much more likely is OPPORTUNITY than NO OPPORTUNITY? | 1 0 0 0 0 5 0
|          |          |       |              |                | 1 0 1 0 0 0 9 1 |
|          |          |       |              |                | 1 0 2 0 0 0 0 0 |
|          |          |       |              |                | 1 0 3 0 0 0 0 1 0 9
<p>|          |          |       |              |                | 1 0 4 0 0 0 5 5 5 |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>NO Benefit</td>
<td>and</td>
<td>SOME Cost</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>MAJOR Benefit</td>
<td>and</td>
<td>NO Cost</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>N</td>
<td>F</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>SOME Benefit</td>
<td>and</td>
<td>NO Cost</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>NO Benefit</td>
<td>and</td>
<td>NO Cost</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

Table B-8 CPT for Capability

<table>
<thead>
<tr>
<th>CAPABILITY</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACCESS</td>
<td>RESOURCES</td>
<td>Centrality</td>
<td>Influence</td>
<td>CAPABILITY</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(23)</td>
</tr>
<tr>
<td>I D</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>V a l</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>ACCESS and Resources and Below Average</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bel ow Avg e a g e</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>I N F</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>NO Access and Resources and Below Average</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bel ow Avg e a g e</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>I N F</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>Access and N O Access and NO Resources and Bel ow Average and Bel ow Average</td>
<td>2</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>N O Access and NO Resources and Bel ow Average and Bel ow Average</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Access and Resources and Bel ow Average and Average</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>N O Access and Resources and Bel ow Average and Average</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>Access and NO Resources and Bel ow Average and Average</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>N O Access and NO Resources and Bel ow Average and Average</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Access and Resources and Bel ow Average and Above Average</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>N O Access and Resources and Bel ow Average and Above Average</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>Access and NO Resources and Bel ow Average and Above Average</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>N a NO a Bel a Above 3</td>
<td>1</td>
</tr>
<tr>
<td>O Access</td>
<td>n d Re sources</td>
<td>n d ow Ave rage</td>
<td>n d ve Ave rage</td>
<td>more likely is NO CAPABILITY than CAPABILITY?</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1 1 2 1 Access and Resources and Average and Below Average 1 1</td>
<td>How much more likely is CAPABILITY than CAPABILITY?</td>
<td>IN F</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2 1 2 1 NO Access and Resources and Average and Below Average 1 1</td>
<td>How much more likely is CAPABILITY than CAPABILITY?</td>
<td>IN F</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1 2 2 1 Access and Resources and Average and Below Average 1 1</td>
<td>How much more likely is CAPABILITY than CAPABILITY?</td>
<td>IN F</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2 2 2 1 NO Access and Resources and Average and Below Average 1 1</td>
<td>How much more likely is CAPABILITY than CAPABILITY?</td>
<td>IN F</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1 1 2 2 Access and Resources and Average and Average 1 1</td>
<td>How much more likely is CAPABILITY than CAPABILITY?</td>
<td>IN F</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2 1 2 2 NO Access and Resources and Average and Average 1 1</td>
<td>How much more likely is CAPABILITY than CAPABILITY?</td>
<td>IN F</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1 2 2 2 Access and Resources and Average and Average 1 1</td>
<td>How much more likely is CAPABILITY than CAPABILITY?</td>
<td>IN F</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2 2 2 2 NO Access and Resources and Average and Average 1 1</td>
<td>How much more likely is CAPABILITY than CAPABILITY?</td>
<td>IN F</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1 1 2 3 Access and Resources and Average and Above Average 5 1</td>
<td>How much more likely is NO CAPABILITY than CAPABILITY?</td>
<td>IN F</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>No.</td>
<td>Access</td>
<td>Resources</td>
<td>Average</td>
<td>Above Average</td>
</tr>
<tr>
<td>-----</td>
<td>---------</td>
<td>-----------</td>
<td>---------</td>
<td>---------------</td>
</tr>
<tr>
<td>1</td>
<td>Acce...</td>
<td>NO Resou...</td>
<td>Ave...</td>
<td>Ave...</td>
</tr>
<tr>
<td>2</td>
<td>NO Acce...</td>
<td>NO Resou...</td>
<td>Ave...</td>
<td>Ave...</td>
</tr>
<tr>
<td>1</td>
<td>Acce...</td>
<td>Resou...</td>
<td>Abo...</td>
<td>Bel...</td>
</tr>
<tr>
<td>2</td>
<td>NO Acce...</td>
<td>Resou...</td>
<td>Abo...</td>
<td>Bel...</td>
</tr>
<tr>
<td>1</td>
<td>Acce...</td>
<td>Resources</td>
<td>Abo...</td>
<td>Bel...</td>
</tr>
<tr>
<td>2</td>
<td>NO Acce...</td>
<td>Resources</td>
<td>Abo...</td>
<td>Bel...</td>
</tr>
<tr>
<td>1</td>
<td>Acce...</td>
<td>Resources</td>
<td>Abo...</td>
<td>Ave...</td>
</tr>
<tr>
<td>2</td>
<td>NO Acce...</td>
<td>Resources</td>
<td>Abo...</td>
<td>Ave...</td>
</tr>
</tbody>
</table>
Table B-9 CPT for Threat

<table>
<thead>
<tr>
<th>THREAT</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO Access</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NO Resources</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Above Average</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Ac</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ce ss rag e</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Ac</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ce ss rag e</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Ac</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ce ss rag e</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Ac</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ce ss rag e</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

How much more likely is NO CAPABILITY than CAPABILITY?

IN 0
<table>
<thead>
<tr>
<th>Variable</th>
<th>INTENT</th>
<th>OPPORTUNITY</th>
<th>CAPABILITY</th>
<th>THREAT</th>
<th>NO THREAT</th>
<th>INF</th>
<th>Val</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 1 1</td>
<td>Int</td>
<td>Oppportunity</td>
<td>Capability</td>
<td>1</td>
<td>2</td>
<td>INF</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>NO Int</td>
<td>Oppportunity</td>
<td>Capability</td>
<td>2</td>
<td>1</td>
<td>INF</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1 2 1</td>
<td>Int</td>
<td>NO Opportun</td>
<td>y</td>
<td>2</td>
<td>1</td>
<td>INF</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>NO Int</td>
<td>NO Opportun</td>
<td>y</td>
<td>2</td>
<td>1</td>
<td>INF</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1 1 2</td>
<td>Int</td>
<td>Oppportunity</td>
<td>NO Capabili</td>
<td>2</td>
<td>1</td>
<td>INF</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>NO Int</td>
<td>Oppportunity</td>
<td>NO Capabili</td>
<td>2</td>
<td>1</td>
<td>INF</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1 2 2</td>
<td>Int</td>
<td>NO Opportun</td>
<td>y</td>
<td>2</td>
<td>1</td>
<td>INF</td>
<td>6</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>NO Int</td>
<td>NO Opportun</td>
<td>y</td>
<td>2</td>
<td>1</td>
<td>INF</td>
<td>7</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table B-10 CPT for Influence
<table>
<thead>
<tr>
<th>NCE</th>
<th>ID</th>
<th>Measures</th>
<th>Influent</th>
<th>Non-Influent</th>
<th>N/A</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>CR &amp; PR</td>
<td>Below Average</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>11 Decreasing &amp; Below Average</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21 Stable &amp; Below Average</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31 Increasing &amp; Below Average</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 Decreasing &amp; Above Average</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22 Stable &amp; Above Average</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32 Increasing &amp; Above Average</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table B-11 CPT for Content

<table>
<thead>
<tr>
<th>Variable</th>
<th>Content</th>
<th>CR</th>
<th>CCR</th>
<th>IRRELEVANT</th>
<th>RELEVANT</th>
<th>N/A</th>
<th>INF</th>
<th>F</th>
<th>0</th>
<th>9</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>INF</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td>How much more likely is IRRELEVANT than RELEVANT?</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>INF</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td></td>
<td>How much more likely is IRRELEVANT than RELEVANT?</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>INF</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td></td>
<td>How much more likely is IRRELEVANT than RELEVANT?</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>INF</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td>How much more likely is IRRELEVANT than RELEVANT?</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>INF</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ID</td>
<td>Stable and Average</td>
<td>Ave rage</td>
<td>Centrality</td>
<td>CPT for Centrality</td>
<td>ID</td>
<td>N</td>
<td>F</td>
<td>Sum</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>--------------------</td>
<td>----------</td>
<td>------------</td>
<td>-------------------</td>
<td>----</td>
<td>---</td>
<td>---</td>
<td>-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Stable</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Increasing</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Decreasing</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Stable</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Increasing</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table B-12 CPT for Centrality
<table>
<thead>
<tr>
<th>3</th>
<th>1</th>
<th>Increa sin g</th>
<th>Below Ave rag e</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>How much more likely is Abnormal than Normal?</th>
<th>5</th>
<th>0</th>
<th>How much more likely is Normal than N/A?</th>
<th>I</th>
<th>N</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>Decre asi ng</td>
<td>Ave rag e</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>How much more likely is Abnormal than Normal?</td>
<td>2</td>
<td>0</td>
<td>How much more likely is Normal than N/A?</td>
<td>I</td>
<td>N</td>
<td>F</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Stable</td>
<td>Ave rag e</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>How much more likely is Abnormal than Normal?</td>
<td>1</td>
<td>0</td>
<td>How much more likely is Normal than N/A?</td>
<td>I</td>
<td>N</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>Increasin g</td>
<td>Ave rag e</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>How much more likely is Abnormal than Normal?</td>
<td>2</td>
<td>0</td>
<td>How much more likely is Normal than N/A?</td>
<td>I</td>
<td>N</td>
<td>F</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>Decreasi ng</td>
<td>Above Ave rag e</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>How much more likely is Abnormal than Normal?</td>
<td>3</td>
<td>0</td>
<td>How much more likely is Normal than N/A?</td>
<td>I</td>
<td>N</td>
<td>F</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>Stable</td>
<td>Above Ave rag e</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>How much more likely is Abnormal than Normal?</td>
<td>8</td>
<td>0</td>
<td>How much more likely is Normal than N/A?</td>
<td>I</td>
<td>N</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>Increasin g</td>
<td>Above Ave rag e</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>How much more likely is Abnormal than Normal?</td>
<td>1</td>
<td>0</td>
<td>How much more likely is Normal than N/A?</td>
<td>I</td>
<td>N</td>
<td>F</td>
</tr>
</tbody>
</table>