International Journal of Innovative
Computing, Information and Control
Volume **9**, Number **9**, September **2013**

ICIC International ©2013 ISSN 1349-4198
pp. 3497–3513

A DATA MINING APPROACH TO ANALYZING STUDENT-PEER RELATIONSHIPS FROM COMMUNICATION HISTORY RECORDS

Yang-Sae Moon, Hun-Young Choi, Jinho Kim and Mi-Jung Choi*

Department of Computer Science
Kangwon National University
192-1 Hyoja2-Dong, Chunchon, Kangwon 200-701, Republic of Korea
{ ysmoon; hychoi; jhkim }@kangwon.ac.kr
*Corresponding author: mjchoi@kangwon.ac.kr

Received July 2012; revised January 2013

Abstract. *In recent years, bullied students and delinquent groups in teenagers cause many serious social problems. In this paper we propose a novel approach that analyzes peer relationships among students more objectively. As the data for objective analysis, we use communication history records that are collected from various communication tools such as telephones, e-mails, short messages, and messengers. We use a simple intuition that communication history records implicitly contain peer relationship information, and we adopt data mining techniques for the more systematic analysis. Our key contribution is to propose a novel data mining-based approach to identifying the potentially bullied students and potentially delinquent groups based on the objective data of communication history records. The proposed method consists of the following steps. First, we formally define the notion of degree of familiarity between students and present mathematical formulas that compute the degree based on communication history records. We here use the intuition that the degree of familiarity from student x to student y becomes higher as the number of communications from x to y increases. Second, using the degree of familiarity we find out the students who are potentially bullied. This procedure is based on the assumption that a bullied student may have a very small number of communication history records issued from other students. Third, we adopt a clustering technique, one of the most representative data mining techniques, to identify meaningful student groups. To use the clustering technique, we first define the measure of similarity between friends based on the degree of familiarity, and we then perform clustering using the similarity measure. Last, to show the practical use of the proposed method, we have implemented the method and interpreted the meaning of experimental results. Overall, we believe that our research result provides a useful framework that analyzes peer relationships among students more objectively and more systematically.*

Keywords: Data mining, Peer relationship, Clustering, Communication history records, Bullied students, Delinquent groups

1. **Introduction.** In recent years, bullied students [19, 22, 26] and delinquent groups [27] in teenagers cause many serious social problems [17, 19, 23]. First, the number of bullied students, who are harassed by or isolated from many other students, is increasing rapidly among teenagers [22, 26, 27]. The problem of bullied/bulling students goes beyond simple harassment and causes serious social problems such as suicide, murder, and family disintegration. Second, the number of delinquent groups such as "Iljinhoe", which has been a notorious delinquent student group in Korea [13], is also increasing due to infelicitous or bad relationships among students. These delinquent groups cause a lot of serious deviations such as using violence and committing blackmail [23]. Basically, we think that bullied students and delinquent groups are caused by infelicitous or bad relationships among students. Therefore, we need to provide an automated solution that

analyzes whole relationships among whole students rather than individual relationships between individual students.

Finding student-peer relationships is very important for providing appropriate student guidance. If we can find those relationships correctly, we may predict some serious deviations and prevent them in advance. By finding objective peer relationships among students systematically, we can recommend that teachers and parents give much more attention and special guidance to the students who may have potential problems in peer relationships.

In this paper we propose a novel approach that analyzes peer relationships among students more objectively. The traditional methods studied in sociology to analyze peer relationships can be divided into two categories [18]. The first one is an observer-oriented approach, where an advisor consults students about their personal affairs or manages history records of students' life. The second one is a student-oriented approach, where students ask for an interview with an advisor and give their information to an advisor. These traditional approaches, however, may find out incorrect peer relationships since they strongly depend on observers' or students' subjective decisions or opinions. Also, the approaches may figure out the distorted relationships among students since they are too simple and straightforward. To solve these problems, we suggest two requirements: (1) the data used for analyzing peer relationships should be objective ones that can be obtained from students' daily life, and (2) the analysis method using the objective data should be the more systematic one. To satisfy these requirements, we use communication history records as the data for objective analysis, and we adopt a data mining technique as the method for systematic analysis.

The analysis method to be proposed is based on *communication history records*. In general, we use various communication tools such as postal mails, telephones, e-mails, short messages, and messengers (including facebook and twitter) to communicate with each other. Among these communication tools, e-mails, short messages, and messengers, which are deployed recently, usually store communication history records generated in communication processes. We here note that communication history records may contain human relationship information implicitly. Simply speaking, the more communications try, the closer relationship has. Also, the communication history records are generated automatically in communication processes, so we can use the records as the objective data to analyze peer-student relationships.

In this paper we propose a systematic approach that identifies bullied students and delinquent groups by using communication history records. In other words, the key contribution of the paper is to propose a novel data mining approach that systematically identifies the isolated students who are potentially bullied and the meaningful student groups that are potentially delinquent by using the communication history records, which have the relatively objective property. The proposed approach consists of the following steps. First, we define the notion of *degree of familiarity* between students and present mathematical formulas that compute the degree based on communication history records. We here use a simple intuition that the degree of familiarity from student x to student y becomes higher as the number of communications from x to y increases. That is, we define the degree of familiarity from student x to student y as how many communications have been tried from x to y. Thus, we can say that the degree of familiarity from x to y is a quantitative criterion that represents how much x likes y as a friend. Second, by using the degree of familiarity between students, we identify the students who might be potentially bullied. This procedure is based on the assumption that a bullied student may have a very small number of communication history records issued from other students. That is, if a student has very low degrees of familiarity from other students, we regard the student

as a bullied one with high confidence. Third, we adopt a clustering technique [6, 7, 8, 15], one of the most representative data mining techniques [9], to identify meaningful student groups using the degree of familiarity. Since the clustering technique is used to distinguish clusters, i.e., groups whose members are similar with each other, we can use the clustering technique to extract meaningful student groups whose members have the higher degree of similarity between each other. To use the clustering technique, we define the notion of *similarity* between friends based on the degree of familiarity. Intuitively speaking, the similarity between two students becomes higher as the number of their common friends increases. As the clustering technique, we use the ROCK algorithm [6, 7] since it has been known as an efficient similarity-based clustering technique.

To show the practical use of our method, we have implemented the proposed method and interpreted the meaning of experimental results. The communication history records used in analysis are collected through a survey. By using the survey data, we first compute the degrees of familiarity for every pair of students. Based on the degrees of familiarity, we then identify the students who are potentially bullied and interpret the meaning of the results. By using the clustering technique, we also extract meaningful student groups and interpret the meaning of the groups. Our method computes the degree of familiarity that is implicitly contained in communication history records, the relatively objective data, and uses the degree to identify bullied students and delinquent groups. Hence, if using our method to analyze peer relationships among students, teachers or parents may recognize bullied students and delinquent groups at an early stage and give a good direction to them in a more objective manner. Our research result can also be used for the recent social network analysis [14]. In particular, the data mining-based peer relationship approach can be applied to anomaly detection and association group analysis in large-scale social networks [16].

The remainder of this paper is organized as follows. In Section 2, we introduce related work on peer relationships among students. In Section 3, we formally define the notion of degree of familiarity between students based on communication history records and present mathematical formulas of computing the degree. In Sections 4 and 5, using the degree of familiarity we propose a systematic approach to identifying bullied students and delinquent groups. In Section 6, we present the implementation and experimental results. In Section 7, we finally conclude and summarize the paper.

2. **Related Work.**

In this section we review the traditional methods to analyzing human or peer relationships in sociology. Sociology is a science which tries to scientifically identify relationships between individuals of a community [23]. Here, a community consists of individuals, and the individuals form a variety of horizontal or vertical relationships in a very complicated manner. In sociology, researchers have focused on major causes of forming complex social relationships [1, 21] and tried to reveal mutual relationships in the causes. Many investigation methodologies such as conflict theory, symbolic interactionism, and structural-functionalism have been studied for human relationship analysis as one of social relationship analysis [2, 10, 23].

The analysis methods which are frequently used in sociology to analyze the relationships between individuals, i.e., human relationships or peer relationships, can be classified into two categories [2, 10]:

- *Interaction analysis*: By analyzing interaction patterns happened among members in a community, interaction analysis identifies relationships between individuals. Bales's "system of categories used in observation and their major relations" and Comel's class observation method are widely used as the representative ones.

- *Sociometry*: By evaluating degrees of attraction, exclusion, and indifference between individuals, sociometry finds each individual's position in a community and analyzes relationships between individuals.

However, since the methods in sociology make a conclusion basically through group members' reports or supervisors' observations, they may cause a lot of misjudgment or misunderstanding in analyzing relationships between individuals [2, 10, 24, 25]. First, wrong conclusions would occur due to inaccurate observations or abnormal subjective views since the sociological methods derive conclusions based on personal observations or subjective views. Second, the methods in sociology may commit a serious mistake that derives an over-generalized conclusion based on a small number of observation cases. Third, since the sociological methods are based on observers' or supervisors' selective data or experience, they may derive an exaggerated conclusion for whole members rather than a small number of members. Fourth, the sociological methods may also make a wrong conclusion derived from the irrational deductions that consequents are irrationally concluded from precedents.

As we have explained above, the traditional methods in sociology have a serious drawback of being hard to ensure objectivity since they have the inherited problem in data collection processes. That is, the input data itself may lack objectivity since the collection of data depends on observers' subjective decisions, individuals' direct reports, and unrelated persons' opinions. Therefore, we need to collect the data from the more objective information sources that can be obtained from individuals' daily life, and we try to use the objective data in analyzing relationships among individuals. To accomplish this purpose, we try to find the degree of familiarity between students from communication history records which are automatically accumulated in students' daily life and propose a data mining approach that uses the degree in analyzing peer relationships among students. Our approach quite differs from sociological methods in terms of data types and base techniques. Also, to our best knowledge, this is the first attempt to use a data mining approach in analyzing students' peer relationships, i.e., identifying bullied students and delinquent groups.

The proposed approach has two important advantages in the viewpoint of effectiveness and efficiency. First, we can find its effectiveness in preventing various social problems by finding student-peer relationships. For example, finding bullied students may prevent the isolated students from leaving school or even killing themselves, and finding delinquent groups may prevent the undesirable student groups from serious deviations such as using violence and committing blackmail. Second, we can find the efficiency of the approach in the computerized analysis method, which is very fast, objective, and automated. More importantly, the proposed approach has an incremental update property. That is, as time goes by, the peer relationships will be changed dynamically. We here note that the proposed approach can reflect these changes incrementally and promptly by simply re-considering the recently added communication history records.

3. Degree of Familiarity from Communication History Records.

Peer relationship is a qualitative measure that represents the degree of close friendship between students. In this section, we propose a systematic approach that quantitatively computes the degree using communication history records, which are the relatively objective data obtained from students' daily life.

In this section, we derive *wdof*(), the weighted degree of familiarity, which will be exploited in identifying bullied students and delinquent groups in Sections 4 and 5. Figure 1 shows how we derive it from communication history records. As shown in the figure, we first explain types, structures, and notations of communication tools and their history

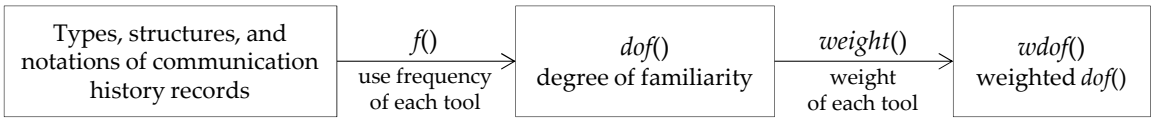

FIGURE 1. A procedure of deriving the weighted degree of familiarity in Section 3

TABLE 1. Examples of communication history records

Communication tools	Configuration of each history record
(celluar) phone	date, outgoing number, incoming number, communication interval
short message	date, outgoing number, incoming number, message
messenger	date, time, sender, receiver, message
e-mail	sender, receiver, data, time, subject, content

records. We next define the use frequency of each tool, denoted as $f()$, and we then present an intuitive equation of computing *dof*(), the degree of familiarity. We finally propose a formal method of computing *wdof*() with the concept of weight of each communication tool.

3.1. **Communication history records.** Recent advances in science and computers enable people to use various communication tools to have conversation with each other. Examples of these communication tools include telephones, e-mails, short messages, and messengers. Recently, facebook and twitter also provide popular messenger functions. We note that most of these communication tools store communication history records that are automatically generated in communication processes. Here, each communication history record is generally composed of sender (or originator) information, receiver (or terminator) information, communication time, etc. We note that these communication history records may contain human relationship information implicitly, i.e., we may assume that the more communications between two people occur, the closer relationship has. Thus, by applying this intuition to students' communication history records, we are able to analyze student-peer relationships. In general, the communication history records, which are generated in communication processes, are stored automatically by computers. Therefore, we may guarantee objectivity of the stored history records and use those records as the objective data to analyze peer relationships.

Most of communication companies store the communication history records for their own purpose. Actually, we are able to obtain our communication history records through formal and legal request to the communication company. Each history record contains sender, receiver, and communication date & time [5]. Table 1 shows examples of communication history records for representative communication tools. As shown in Table 1, most of communication history records contain the information of a sender (or an originator) and a receiver (or a terminator). In this paper, we use this limited information of senders and receivers to analyze peer relationships among students. That is, we perform the analysis using the minimum information of history records rather than using the whole private communication contents. To analyze peer relationships among students based on communication history records, we summarize in Table 2 the notation to be used throughout the paper.

3.2. **Degree of familiarity.** In this section, we formally define the notion of degree of familiarity between students, and we propose a formal method that computes the familiarity degree mathematically. We first define the degree of familiarity based on

TABLE 2. Summary of notation

Symbols	Definitions
x, y, z	Students to be analyzed
S	A set of students to be analyzed (i.e., $x \in S$, $y \in S$, $z \in S$)
m	Number of communication tools
$f_i(x, y)$	Number of communications from x to y using the i-th communication tool f_i $(1 \leq i \leq m)$

TABLE 3. An example of the number of communications that student x tried

category	phone ($f_1(\cdot,\cdot)$)	e-mail ($f_2(\cdot,\cdot)$)	short message ($f_3(\cdot,\cdot)$)
# of whole communications	100	50	200
# of communications to student y	20	15	50

the intuition that "the more communications tried from student x to student y, the closer friendship from x to y". According to this intuition, we then define the degree of familiarity as follows.

Definition 3.1. *Given communication history records, the degree of familiarity from student x to student y is defined as the ratio of "the number of all communications that student x tried" to "the number of communications that student x tried to student y". That is, we mathematically define $dof(x, y)$, degree of familiarity from x to y, as Equation (1):*

$$dof(x,y) = \frac{\#\text{ of communications that student } x \text{ tried student } y}{\#\text{ of all communications that student } x \text{ tried}} = \frac{\sum_{i=1}^{m} f_i(x,y)}{\sum_{\text{for all } z \in S} \left(\sum_{i=1}^{m} f_i(x,z)\right)} \tag{1}$$

According to Definition 3.1, $dof(x, y)$ becomes higher as the number of communications from x to y increases. In general, a student who formed a close friendship with another student may attempt frequent communications to that student. Thus, we think the degree in Equation (1) of Definition 3.1 is a reasonable and intuitional measure that reflects students' daily life.

Example 3.1. *Suppose Table 3 shows the number of communications that student x tried. As shown in Table 3, student x uses three communication tools, phone, e-mail, and short messages. In this example, we can compute the degree of familiarity from x to y as 0.24 by using Definition 3.1.*

$$dof(x,y) = \frac{\sum_{i=1}^{3} f_i(x,y)}{\sum_{\text{for all } z \in S} \left(\sum_{i=1}^{3} f_i(x,z)\right)} = \frac{\sum_{i=1}^{3} f_i(x,y)}{\sum_{i=1}^{3} \left(\sum_{\text{for all } z \in S} f_i(x,z)\right)} = \frac{20+15+50}{100+50+200} = \frac{85}{350} \cong 0.24. \tag{2}$$

However, personal preferences may differ in using communication tools, and thus, we need to consider the preferences in computing the degree of familiarity. That is, when computing the degree of familiarity, we need to give larger weights to one's preferred communication tools. To consider the preferences of individual students, we compute the weight of a specific communication tool as the ratio of the number of all communications to the number of communications using that tool and apply the weight to compute the

degree of familiarity. We now formally define the weight of a communication tool as follows.

Definition 3.2. *Given communication history records, the weight $weight(x,k)$ of the k-th communication tool for student x is defined as the multiplication of the number of communication tools and the ratio of the k-th tool against all tools in the number of communications that student x tried. That is, we mathematically define $weight(x,y)$ as Equation (3):*

$$\begin{aligned} weight(x,k) &= \#\text{ of communication tools} \\ &\quad \cdot \frac{\#\text{ of communications of using } k^{\text{th}} \text{ tool by student } x}{\#\text{ of all communications that student } x \text{ tried}} \\ &= m \cdot \frac{\sum_{\text{for all } y \in S} f_k(x,y)}{\sum_{i=1}^{m} (\sum_{\text{for all } y \in S} f_i(x,y))} \end{aligned} \tag{3}$$

By using the weight in Definition 3.2, we can redefine the degree of familiarity presented in Definition 3.1. That is, if we want to consider the preference of each tool, we may evolve the number of communications from student x to student y from $\sum_{i=1}^{m} f_i(x,y)$ to $\sum_{i=1}^{m}(weight(x,k) \cdot f_i(x,y))$. Thus, we call the number of communications that can be computed using the weight as the *weighted number of communications*. By using the weighted number of communications, we also define the weighted degree of familiarity as follows.

Definition 3.3. *Given communication history records, the weighted degree of familiarity from student x to student y is defined as the ratio of "the number of all communications that student x tried" to "the weighted number of communications that student x tried to student y". That is, we mathematically define the weighted degree of familiarity from x to y, $wdof(x,y)$, as Equation (4):*

$$\begin{aligned} wdof(x,y) &= \frac{\text{the } \textit{weighted} \text{ number of communications that student } x \text{ tried to student } y}{\#\text{ of all communications that student } x \text{ tried}} \\ &= \frac{\sum_{i=1}^{m}(weight(x,i) \cdot f_i(x,y))}{\sum_{\text{for all } z \in S} (\sum_{i=1}^{m} f_i(x,z))} \end{aligned} \tag{4}$$

Example 3.2. *As in Example 3.1, suppose Table 3 in Example 3.1 shows the number of communications that student x tried. Then, we can first compute the weight of each communication tool for student x as follows:*

$$\begin{aligned} weight(x,1) &= 3 \cdot \frac{\sum_{\text{for all } y \in S} f_1(x,y)}{\sum_{i=1}^{3} (\sum_{\text{for all } y \in S} f_i(x,y))} = 3 \cdot \frac{100}{100+50+200} = \frac{300}{350} \cong 0.86 \\ weight(x,2) &= 3 \cdot \frac{\sum_{\text{for all } y \in S} f_2(x,y)}{\sum_{i=1}^{3} (\sum_{\text{for all } y \in S} f_i(x,y))} = \frac{150}{350} \cong 0.43 \\ weight(x,3) &= 3 \cdot \frac{\sum_{\text{for all } y \in S} f_3(x,y)}{\sum_{i=1}^{3} (\sum_{\text{for all } y \in S} f_i(x,y))} = \frac{600}{350} \cong 1.71 \end{aligned}$$

According to the results above, in case of student x, we give the weight 0.86 to phones (f_1) to treat one communication try as 0.86 tries, 0.43 to e-mail (f_2), and 1.71 to short messages (f_3). That is, we give the largest weight to short messages which student x most frequently uses. Based on these weight values, we can compute the weighted degree of

familiarity from student x to student y as follows:

$$wdof(x,y) = \frac{\sum_{i=1}^{3}(weight(x,i)\cdot f_i(x,y))}{\sum_{\text{for all } z\in S}(\sum_{i=1}^{3} f_i(x,z))} = \frac{0.86\cdot 20 + 0.43\cdot 15 + 1.71\cdot 50}{350} \cong \frac{109.2}{350} \cong 0.3 \tag{5}$$

The difference between the degree of familiarity in Definition 3.1 and the weighted degree of familiarity in Definition 3.3 is in whether we consider personal preferences of communication tools or not. In particular, we can see there is a relatively large difference between two degrees in Examples 3.1 and 3.2. It is because student x frequently uses short messages compared with other communication tools, phones and e-mails. In the examples, the ratio of using short messages from student x to student y is relatively high compared with those of phones and e-mails, and thus, the weighted degree of familiarity from x to y becomes larger than the original degree of familiarity. That is, the degree of familiarity in Definition 3.1 uses the number of communications without consideration of personal preferences of communication tools, while the weighted degree of familiarity in Definition 3.3 uses the weighted number of communications with consideration of personal preferences. In this paper we use the weighted degree of familiarity rather than the original degree of familiarity to reflect personal preferences of communication tools in analyzing peer relationships among students. Hereafter, we use the weighted degree of familiarity and the degree of familiarity interchangeablly unless confusion occurs.

Figure 2 shows an overall process that computes the degrees of familiarity from the communication history records. As shown in the figure, the input to the computation algorithm is a set of communication history records; the output a set of the degrees of familiarity between students. Here, the degree of familiarity between students is computed using the weight in Definition 3.2 and the weighted degree of familiarity in Definition 3.3. Algorithm 1 shows *ComputeWdof*() that computes the degrees of familiarity between students. The input to the algorithm is a set of communication history records $f_i(x,y)$ automatically accumulated in communication processes. The algorithm consists of a series of steps that compute each degree of familiarity from x to y for every pair of students x and y. In Steps (1) to (3), we first compute the weight of each communication tool for every student x by using Equation (3) of Definition 3.2. In Steps (4) to (6), we then compute each degree of familiarity from x to y for every pair of students x and y by using Equation (4) of Definition 3.3.

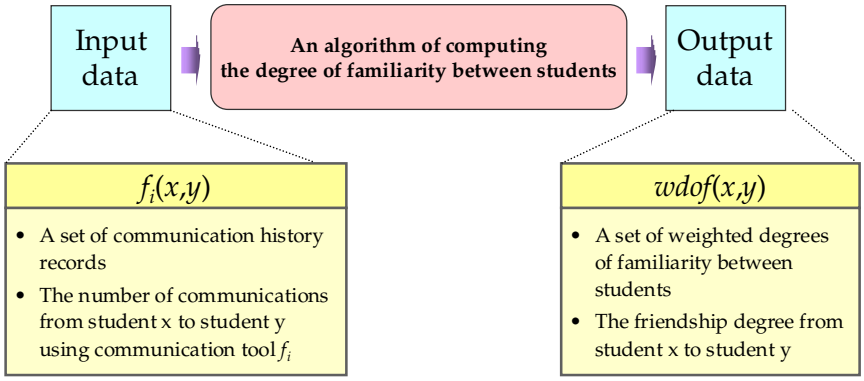

FIGURE 2. An overall process of computing degrees of familiarity between students

Algorithm 1 *ComputeWdof*(A set of $f_i(x, y)$'s)

1: **for each** x **in** S **do**
2: **for each** k **in** $[1, m]$ **do**
3: $weight(x, k) = m \cdot \frac{\sum_{\text{for all } y \in S} f_k(x,y)}{\sum_{i=1}^{m}(\sum_{\text{for all } y \in S} f_i(x,y))}$; // weight (preference) of tool k for student x
4: **end for**
5: **end for**
6: **for each** x **in** S **do**
7: **for each** y **in** $S\,(y \neq x)$ **do**
8: $wdof(x, y) = \frac{\sum_{i=1}^{m}(weight(x,i) \cdot f_i(x,y))}{\sum_{\text{for all } z \in S}(\sum_{i=1}^{m} f_i(x,z))}$; // weighted degree of familiarity from x to y
9: **end for**
10: **end for**

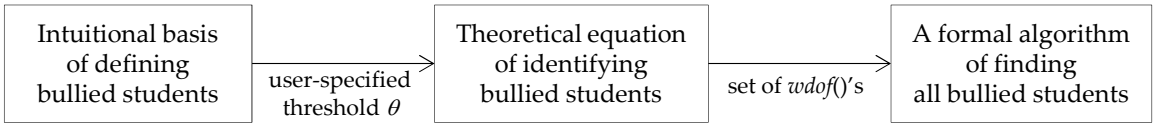

FIGURE 3. A procedure of deriving a formal algorithm of identifying bullied students in Section 4

4. **Data Mining Approach to Identifying Bullied Students.** By using the degree of familiarity explained in Section 3, we propose a novel method of identifying the bullied students who are potentially harassed or isolated and the meaningful student groups that are potentially delinquent. In this section, we present an intuitional basis that finds out bullied students using the sum over degrees of familiarity and propose an efficient algorithm based on the intuition. In the next section, we adopt a clustering technique to identify delinquent groups using the degree of familiarity and propose another efficient algorithm based on the clustering technique.

In this section, we eventually obtain a formal algorithm that identifies bullied students from *wdof*()'s of Section 3. Figure 3 shows the procedure of deriving the algorithm. As shown in the figure, we first present the intuitional concept of how we define the bullied students. We then derive a theoretical equation that identifies bullied students by introducing the user-specified tolerance θ. We finally propose a formal algorithm that finds all bullied students from the weighted degrees of familiarity that are computed from communication history records in Section 3.

In general, we may regard a student who is isolated from many other students as a bullied one. That is, we may think that a student is potentially bullied if many other students do not consider him/her as a friend. Based on this intuition, we present the following assumption to identify the bullied students using the degree of familiarity.

Assumption 1: When computing the degree of familiarity between students based on communication history records, we may regard a student whose sum over degrees of familiarity is very low as a bullied one with high confidence.

As we explained in Section 3, the degree of familiarity from student x to student y is defined as how many communications have been tried from x to y. Hence, if using Assumption 1, we may identify a student who has a very small number of communications from other students as a bullied one. We think that this assumption is relatively reasonable and objective to find out bullied students since the bullied students might have a very little number of friends.

Based on Assumption 1, we propose a systematic way of identifying bullied students as follows. We first compute $\sum_{\text{for all } x \in S} wdof(x, y)$ for student y, which is the sum over degrees of familiarity from all the other students to y. We then identify a student whose

sum over degrees is very low as a bullied one. Equation (6) shows a mathematical formula for this determination process.

$$\sum_{\text{for all } x \in S} wdof(x, y) \simeq 0 \tag{6}$$

We here need to represent "near 0" as a quantitative measure in deciding bullied students. As a quantitative approach, we now introduce a threshold θ to formally define bullied students as follows.

Definition 4.1. *If a student's sum over degrees of familiarity from all the other students is less than or equal to the user-specified threshold θ, then the student can be regarded as a bullied student. That is, student y is defined as a bullied student if Equation (7) is satisfied.*

$$\sum_{\text{for all } x \in S} wdof(x, y) \leq \theta \tag{7}$$

Based on Definition 4.1, we are able to identify the bullied student whose sum over the degrees of familiarity is less than or equal to the user-specified threshold θ, which will be given by an analyzer such as a teacher and a supervisor.

Algorithm 2 shows *FindBullied*() that identifies the bullied students based on Definition 4.1. The inputs to the algorithm are a set of *wdof*(x, y)'s, the degrees of familiarity, and the threshold θ. As shown in the algorithm, we first compute the sum over degrees for each student y, and we then decide y as a bullied one if the sum is less than or equal to the given threshold θ. In Step (2), we mark student y as a bullied one since his/her sum over degrees from all the other students is less than or equal to θ. On the other hand, in Step (3) we mark y as a normal one since the sum is greater than θ.

Algorithm 2 *FindBullied*(A set of $wdof(x, y)$'s, Threshold θ)

1: **for each** y **in** S **do**
2: **if** $\sum_{\text{for all } x \in S} wdof(x, y) \leq \theta$ **then** $bullied(y) =$ `True`;
3: **else** $bullied(y) =$ `False`;
4: **end for**

5. **Data Mining Approach to Identifying Delinquent Groups.** We also use the degree of familiarity to extract meaningful student groups from communication history records. However, we cannot directly use the degree of familiarity to identifying the groups since the degree is not adequate to use in existing clustering algorithms. Hence, we introduce the similarity between students as a new measure. Based on the measure, we then adopt a clustering technique [4, 8, 11, 15] to distinguish clusters, i.e., groups whose members are similar with each other. As a clustering algorithm, we use ROCK [6, 7], which is widely used in the case where the similarity between two objects can be obtained in advance.

Figure 4 shows the proposed systematic procedure of identifying meaningful student groups, i.e., delinquent student groups, from *wdof*()'s of Section 3. As shown in the figure, we first obtain *flist*(), which represents a list of friends for a specific student from his/her *wdof*()'s. Using Jacaard index [12], we next define *sim*(), which represents a quantitative similarity between two students. By integrating these two measures, we then propose an algorithm that computes similarity measures for all possible student pairs. We finally adopt ROCK, a representative clustering algorithm, to find meaningful student groups, which might be potentially delinquent student groups.

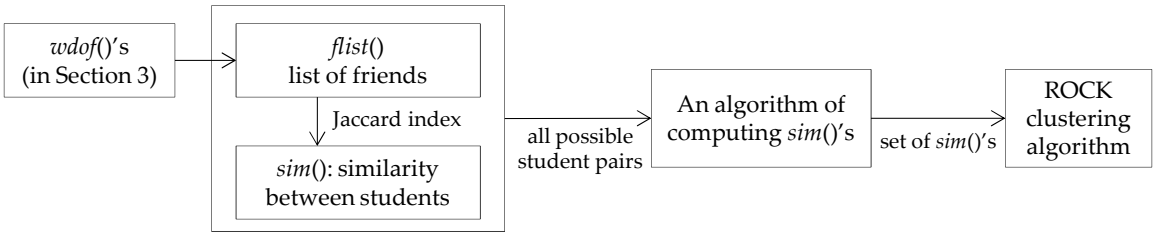

FIGURE 4. The proposed procedure of identifying delinquent student groups in Section 5

To use the algorithm ROCK, we have to define the similarity between two objects, i.e., two students. This is because most of clustering techniques including ROCK perform clustering based on the similarity (or the distance) between two objects. Hence, to identify meaningful student groups using ROCK, we first compute the similarity between two students based on communication history records. By the way, in Section 3 we have already defined the degree of familiarity between students and also proposed an algorithm of computing the degrees from communication history records. Thus, by using the degree of familiarity, we formally define the measure of similarity between two students and propose an algorithm of computing the similarity for ROCK.

Intuitively speaking, the similarity between two students becomes higher as the number of their common friends increases. That is, if two students have a large number of common friends, we can say that they are in close friendship and are similar with each other. By this intuition, we conceptually define the similarity between two students x and y as "how many friends of x are similar with those of y", i.e., "how many common friends x and y have". To compute the similarity between students based on such an intuitional concept, we need to obtain the list of friends for each student first. We include student y into the list $\mathit{flist}(x)$ of friends for student x if the degree of familiarity from x to y is greater than another user-specified threshold δ. Here, the threshold δ will also be given by an analyzer such as a teacher and a supervisor, and its value can be larger than 0. That is, the threshold δ is a quantitative criterion of how often student x tries to communicate with student y in order to regard y as x's friend. To simplify the problem, however, we use 0 as the threshold δ, i.e., we assume the students with whom x tries to communicate at least one time as x's friends. According to this assumption, if $\mathit{wdof}(x, y)$ is greater than $0(=\delta)$, we include student y into $\mathit{flist}(x)$, the list of friends for student x.

By using the list of friends for each student, we define the similarity $\mathit{sim}(x, y)$ between two students x and y as follows.

Definition 5.1. *Given flist(x) and flist(y) that are the lists of friends for students x and y, respectively, the similarity between x and y, sim(x, y), is defined as the ratio of the union of two lists to the intersection of two lists. That is, the similarity between x and y is defined as Equation (8):*

$$\mathit{sim}(x, y) = \frac{|\ \mathit{flist}(x) \cap \mathit{flist}(y)\ |}{|\ \mathit{flist}(x) \cup \mathit{flist}(y)\ |} \tag{8}$$

In brief, the similarity in Equation (8) of Definition 5.1 means the ratio of common friends of two students. That is, the divisor in Equation (8), the union of two lists, represents all the friends of two students; the dividend, the intersection of two lists, the common friends of them. Thus, we may regard Equation (8) as the degree of how many common friends two students have, and we are able to use the similarity in Definition 5.1 as the similarity measure in a clustering algorithm.

Algorithm 3 shows *ComputeSim*() that computes the similarity between each pair of students based on the degree of familiarity. The inputs to the algorithm are a set of *wdof*(x, y)'s and the threshold δ. In Steps (1) to (5), we first compute *flist*(x), the list of friends for each student x, by using the degree of familiarity. In Steps (6) to (10), we then compute the similarity between two students x and y by using the lists of friends for x and y.

Algorithm 3 *ComputeSim*(A set of *wdof*(x, y)'s, Threshold δ)

```
1: for each x in S do
2:   for each y in S (y ≠ x) do
3:     if wdof(x, y) > δ then Include y into flist(x);    // the list of friends for x
4:   end for
5: end for
6: for each x in S do
7:   for each y in S (y ≠ x) do
8:     sim(x, y) = |flist(x) ∪ flist(y)| / |flist(x) ∩ flist(y)|;    // the similarity for all possible pairs of students
9:   end for
10: end for
```

After computing the similarity $sim(x, y)$ for every pair of students using the algorithm *ComputeSim*(), we now perform clustering by using that similarity measure. For this purpose, we assume an intuitive relation between the similarity and the cluster as follows. **Assumption 2**: When computing the similarity between students based on communication history records, we may think the two students who have a large similarity value will be contained in the same cluster (the same student group) with high confidence.

According to Assumption 2, we can identify clusters, i.e., meaningful student groups, by applying the similarity between students to the algorithm ROCK [7][1].

6. **Experiments and Analysis.** In this section we explain the implementation of the proposed method and present real experimental results on high school students. As we mentioned in Section 2, the previous sociological methods for peer relationship analysis are quite different from our data mining-based approach in terms of data types and base techniques. Thus, in this section, we do our best to analyze the experimental results more intuitively and more objectively. First, in Section 6.1, we introduce implementation and experimental environment. Second, in Sections 6.2 and 6.3, we explain the experimental results for bullied students and delinquent groups, respectively.

6.1. **Implementation and experimental environment.** We implemented all the algorithms proposed in Sections 3 to 5. We first implemented the algorithm *ComputeWdof*() proposed in Section 3 to obtain the degree of familiarity between students. We then implemented the algorithm *FindBullied*() for identifying bullied students. We also implemented the algorithm *ComputeSim*() and the clustering algorithm ROCK for extracting delinquent groups. We conducted all the experiments on a Windows 7 server with Intel Core2 Duo 2.53GHz CPU, 2GB RAM, and 500GB hard disk and used Borland Delphi language [3] to implement the algorithms.

We collected the communication history records through a survey. We could obtain the real history records from telecommunication companies only after we got legal agrees of individual students due to "communication secret guard law [5]" and "personal information guard law [20]". In our situation, however, it was difficult to get all the necessary

[1]The algorithm ROCK consists of *ComputeLinks*(), which computes links, and *Cluster*(), which performs clustering. These procedures, however, are not a main focus of the paper, and we omit the details.

TABLE 4. The number and the ratio of communications for each tool

Category	Phone	Short message	Messenger	e-mail
Number of communications	1,993	10,441	28,391	106
Ratio of communications (%)	4.9	25.5	69.4	0.3

agrees from every student, and thus, we collected the history records through a survey[2]. The survey was performed on all the second year high school students of six classes located in a city of Korea. The items surveyed for each student (i.e., for each sender) consist of receivers (or terminators) and the numbers of communications for representative communication tools including cellular phones, short messages, messengers, and e-mails.

The number of students participated in the survey is 143, and Table 4 shows the number and the ratio of communications for each tool. Statistical results in Table 4 show that most of students frequently use messengers and short messages as their major communication tools. On the other hand, the ratio of using phones is comparatively low, and e-mails are hardly used.

6.2. **Analysis for bullied students.** As proposed in Section 4, we regard student y whose sum over degrees of familiarity, $\sum_{\text{for all } x \in S} wdof(x, y)$, is less than or equal to the threshold θ as a bullied student. Thus, in the analysis for bullied students, we perform the experiments by varying the threshold θ and report the bullied students for each θ.

According to the experimental results, the average value of sums over degrees of familiarity for all students is 0.90. The reason why the average is below 1 is that there are a few students who do not make any communication with other students. That is, in case of those students, the sum over degrees of familiarity to other students becomes 0 instead of 1. On the basis of the average 0.90, we can interpret each student's sum over degrees of familiarity as follows: (1) a student whose sum over degrees is less than the average may have a small number of friends who like that student; but (2) a student whose sum is greater than the average may have a large number of friends.

Figure 5 shows changes in the number of students when we vary the threshold θ from 0.1 to 3.0. As shown in the figure, most of students (93.0% of all the students) are evenly distributed between 0.0 and 1.71. Intuitively, we expected that most of students would be concentrated around the average 0.90. According to the results, however, students are evenly distributed over wide ranges. This means that student-peer relationships are formed in a quite complex manner with individuality rather than in a simple manner with common sense. Next, a student whose sum over degrees of familiarity has the lower threshold value such as 0.10 or 0.20 might be identified as a bullied one. Table 5 shows the number of students and their identifiers[3] whose sums over degrees of familiarity are less than or equal to the given threshold. That is, the students listed in the table have high possibility of being identified as the actual bullied students. As shown in the table, the students who are identified as the bullied students when the threshold is 0.20, i.e., who have relatively high possibility of being identified as the actual bullied students, are about 10.5% of all students. Also, we can see that the students who are identified as bullied

[2]It is feasible for teachers or parents to obtain the communication history records (especially, the limited information such as senders and receivers) by taking their students' agrees for the purpose of analyzing student-peer relationships. In this research, however, we focus on an analysis method itself rather than correctness or completeness of the records. Thus, we conclude to collect the records through a survey as the simplest way.

[3]In convenience, each student identifier is represented as three digits: the first digit means the class identifier, and the rest two digits mean each student's own identifier.

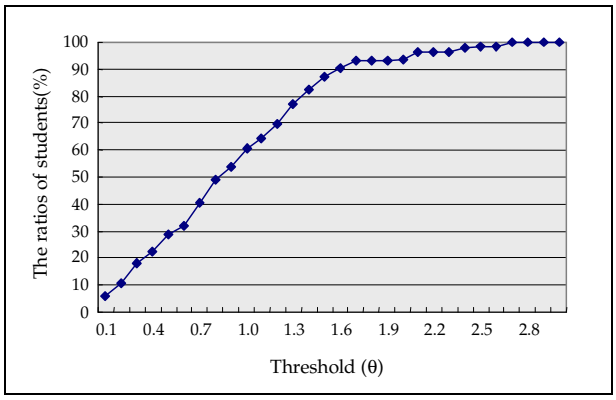

FIGURE 5. Ratios of students when changing the user-specified threshold θ

TABLE 5. The bullied students identified through experiments

Threshold (θ)	# of students	Ratio of students (%)	Student identifiers
0.00	1	0.7	606
0.05	3	1.4	606, 505, 515
0.10	8	5.6	606, 505, 515, 127, 512, 504, 104, 511
0.15	11	7.7	606, 505, 515, 127, 512, 504, 104, 511, 324, 603, 107
0.20	15	10.5	606, 505, 515, 127, 512, 504, 104, 511, 324, 603, 107, 419, 209, 506, 617

students when the threshold is 0.10, i.e., who have very high possibility of being identified as the actual bullied students, are about 5.6% of all students. We now recommend that teachers or parents should give much more attention and special guidance to the students identified as bullied ones.

6.3. **Analysis for delinquent groups.** By implementing the algorithm *ComputeSim*() and the clustering algorithm ROCK, we tried to identify delinquent groups. In the experiment, we fixed the number of groups (clusters) to six. That is, we performed the experiments by setting the number of clusters to be extracted to six which was the same as the number of classes surveyed for the experiments.

We intuitively expected that the pattern of extracted clusters would have some meaningful relationship with the structure of classes since the number of clusters was equal to the number of classes. That is, we expected that the students in the same cluster might be contained in the same class. However, the expectation is not correct. Experimental results show that the students in the same cluster are widely distributed over all the classes. Figure 6 shows six clusters and their student identifiers, and Figure 7 shows a chart of students' distributions over classes for each cluster. As shown in the figures, we can see that the clusters are formed over several classes rather than over one or two specific classes. In particular, we can observe that Cluster 2 contains about 41% of all students, and this confirms the fact that student groups are formed over all the classes rather than within individual classes. This conclusion indicates that it is required for teachers or parents to guide individual students by considering inter-class student relationships as well as intra-class student relationships.

In Figures 6 and 7, Clusters 3, 4, 5, and 6 are small clusters that have a very small number of members. We think these small clusters will be different from the normal clusters, Clusters 1 and 2, that contain most of students. In other words, these small groups may be identified as abnormal groups that are isolated from most of students or as

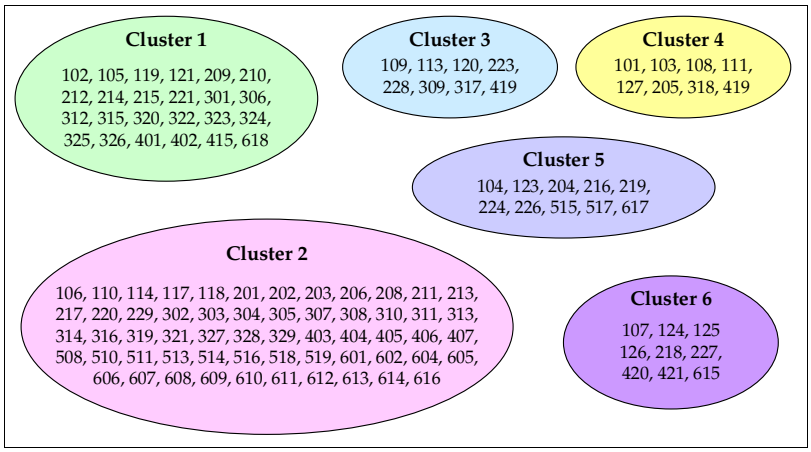

FIGURE 6. Student's distribution over clusters extracted by the clustering algorithm

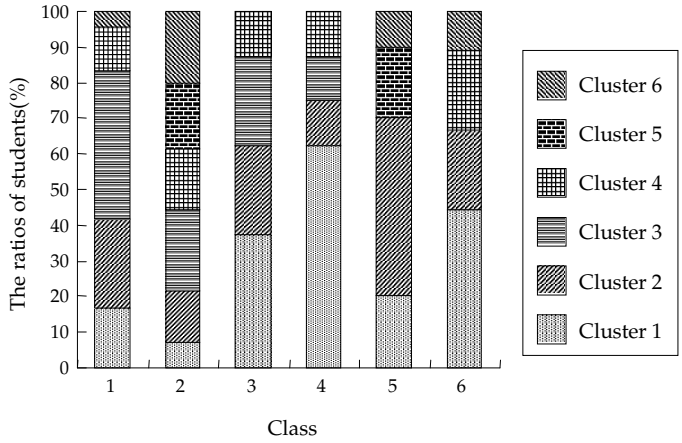

FIGURE 7. Student's distribution over clusters for each class

TABLE 6. Outlier students obtained from clustering analysis

Outlier identifiers (incoming familiarity ≥ 0.5)	Outlier identifiers (incoming familiarity < 0.5)
115(1.51), 116(1.49), 122(0.91), 207(0.87), 222(0.99), 225(0.99), 410(1.67), 503(0.60), 509(1.55)	112(0.36), 409(0.23), 507(0.28), 512(0.06), 603(0.11)

delinquent groups that use violence or commit blackmail. Therefore, we also recommend that teachers or parents should give much more attention to these small student groups.

Table 6 summarizes outliers obtained from the clustering result. That is, outliers of Table 6 mean the students who are not included in any one of six clusters. These outliers have no communication attempt to any other students, and they might think that other students in the experiment are not their true friends. Outlier students of Table 6 are classified into two categories: one for higher incoming familiarity, and the other for lower incoming familiarity. The outliers with higher (≥ 0.5) incoming familiarity might have a few potential friends since some students want to communicate with them. On the other hand, the outliers with lower (< 0.5) incoming familiarity might have no friend since few students want to communicate with them. Therefore, we can highly recommend that teachers or parents should give much more attention to those outlier students having lower incoming familiarity.

7. **Conclusions.** In this paper we proposed an objective and systematic approach to analyzing student-peer relationships. Previous efforts lacked objectivity since they analyzed peer relationships based on personal observations or subjective reports. In contrast, our approach uses communication history records as the objective data to analyze the peer relationships. By using the communication history records, we also proposed a systematic way of identifying bullied students or delinquent groups.

The key contribution of the paper can be summarized as proposing a novel data mining-based approach to identifying the potentially bullied students and potentially delinquent groups based on the objective data of communication history records. The proposed data mining approach was composed of the following steps. First, we formally defined the degree of familiarity and presented mathematical formulas that compute the degree from the communication history records. Second, we proposed a novel approach that identifies bullied students by using the degree of familiarity. To formalize the method, we assumed that a bullied student might have a very small number of communication history records issued from other students, and using this assumption we proposed an intuitive algorithm of identifying the bullied students. Third, we adopted the clustering technique to identify meaningful student groups. To use the clustering technique, we defined the measure of similarity between students and proposed an algorithm of computing the similarity from the degrees of familiarity. Last, we showed the practical use of our analysis method by implementing the method and by interpreting the meaning of experimental results. These results indicate that our research result provides a useful framework that analyzes student-peer relationships more objectively and more systematically.

As further research, we need to validate our method by comparing the experimental results with the actual student-peer relationships. We will also provide user guidance or manuals to use the experimental results in real educational environments. Furthermore, we will try to apply our data mining-based approach to peer relationship analysis in social networks.

Acknowledgement. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012-0005258).

REFERENCES

[1] W. Aalst and M. Song, Mining social networks: Uncovering interaction patterns in business processes, *Proc. of the 2nd Conf. on Business Process Management*, Postdam, Germany, pp.244-260, 2004.
[2] J. Ashford and C. LeCroy, *Human Behavior in the Social Environment: A Multidimensional Perspective*, 4th Edition, Books/Cole, Cengage Learning, 2009.
[3] M. Cantu, *Mastering Borland Delphi 2005*, John Wiley & Sons, 2005.
[4] C.-H. Cheng, J.-R. Chang and L.-Y. Wei, Adaptive-clustering based method to estimate null values in relational databases, *International Journal of Innovative Computing, Information and Control*, vol.7, no.1, pp.223-236, 2011.
[5] *Communication Secret Guard Law*, Initial Legislation in 1993 (Law no.4650), The 17th Revision in 2009 (Law no.9819) (in Korean), 2009.
[6] M. Dutta, A. Mahanta and A. Pujari, QROCK: A quick version of the ROCK algorithm for clustering of categorical data, *Journal of Pattern Recognition Letters*, vol.26, no.15, pp.2364-2373, 2005.
[7] S. Guha, R. Rastogi and K. Shim, ROCK: A robust clustering algorithm for categorical attributes, *Proc. IEEE Int'l Conf. on Data Engineering*, Sydney, Australia, pp.512-521, 1999.
[8] M. Haikidi and M. Vazirgiannis, A density-based cluster validity approach using multi-representatives, *Pattern Recognition Letters*, vol.29, no.6, pp.773-786, 2008.
[9] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, 3rd Edition, Morgan Kaufmann, 2011.
[10] B. Herbert, *Symbolic Interactionism: Perspective and Method*, Prentice-Hall, 1969.

[11] J.-J. Hwang, K.-Y. Whang, Y.-S. Moon and B.-S. Lee, A top-down approach for density-based clustering using multidimensional indexes, *Journal of Systems and Software*, vol.73, no.1, pp.169-180, 2004.
[12] P. Jaccard, Étude comparative de la distribution Florale Dans une portion des Alpes et Des Jura, *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol.37, pp.547-579, 1901.
[13] N.-K. Kim, *Globalization and Regional Integration in Europe and Asia*, Ashgate Publishing Ltd., 2009.
[14] D. Knoke and S. Yang, *Social Network Analysis*, 2nd Edition, Sage Publications, 2008.
[15] H.-P. Kriegel and P. Kroger, Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering, *ACM Transactions on Knowledge Discovery from Data*, vol.3, no.1, 2009.
[16] O. Kwon and Y. Wen, An empirical study of the factors affecting social network service use, *Computers in Human Behavior*, vol.26, no.2, pp.254-263, 2010.
[17] C. Lee, *Preventing Bullying in Schools*, Sage, 2004.
[18] I. Marsh, *Theory and Practices in Sociology*, Pearson Education, 2002.
[19] B. Mary, T. Argus and T. Remley, Bullying and school violence: A proposed prevention program, *NASSP Bulletin*, pp.38-47, 1999.
[20] *Personal Information Guard Law*, Initial Legislation in 1994 (Law no.4743), The 1st Revision in 2011 (Law no.10465) (in Korean), 2011.
[21] J. Saltz, S. Hiltz and M. Turoff, Student social graphs: Visualizing a student's online social network, *Proc. of Conference on Computer Supported Cooperative Work*, ACM CSCW, Chicago, IL, USA, pp.596-599, 2004.
[22] R. Sampson, *Bullying in School*, Problem-Oriented Guides for Police Series, Office of Community Oriented Policing Services, U.S. Department of Justice, 2004.
[23] L. Siegel and B. Wels, *Juvenile Delinquency: Theory, Practice, and Law*, 11th Edition, Wadswroth, Cengage Learning, 2011.
[24] H. Snell et al., *Social Relationships and Peer Support*, Paul H. Brooks Pub. Co., 1999.
[25] J. Wajcman, *Feminism Confronts Technology*, Pennsylvania State Univ. Press, 1991.
[26] J. Wang, R. Iannotti and T. Nansel, School bullying among adolescents in the united states: Physical, verbal, relational, and Cyber, *Journal of Adolescent Health*, vol.45, no.4, pp.368-375, 2009.
[27] M. Warr, Organization and instigation in Deliquent groups, *Criminology*, vol.34, no.1, pp.11-37, 1996.