Identifying key players in a network of child exploitation websites using Principal Component Analysis

F. Movahedi (f.movahedi@gu.ac.ir)
Golestan University
R. Frank
Simon Fraser University

Research Article

Keywords: Social network analysis, Child exploitation, Network disruption, Attack strategy, Web content.

Posted Date: June 15th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1726998/v1

Identifying key players in a network of child exploitation websites using Principal Component Analysis

F. Movahedi[a,b,1], R. Frank[c]

a. Department of Mathematics, Faculty of Sciences, Golestan University, Gorgan, Iran.
b. MoCSSy Program, The IRMACS Centre, Simon Fraser University, Burnaby, British Columbia, Canada.
c. School of Criminology International CyberCrime Research Centre, Simon Fraser University, 8888 University Drive, Burnaby, B.C., Canada. V5A 1S6.

Abstract

One of the main objectives of this study is to help prioritize targets for law enforcement by analyzing online websites hosting child exploitation material and finding key players within. Key players are defined as the websites which display a combination of high connectivity and a lot of hardcore material and would provide the most disruption in a network if they were to be removed. In this study, various strategies based on Principal Component Analysis are presented to identify those nodes that act as the key players in an online child exploitation network. For evaluating the results of these strategies, we consider the results of various attack strategies. The measures for evaluation are the density, clustering coefficient, average path length, diameter and the number of connected components in the resulted network. The results show that the strategies proposed are more successful at reducing all of the outcome measures than existing strategies.

Keywords: Social network analysis, Child exploitation, Network disruption, Attack strategy, Web content.

1. Introduction

The Internet has provided the social, individual, and technological conditions needed for child exploitation to flourish online. In 2009, the United Nations estimated that there were over four million websites containing such content [1] while the UK alone processed 105,047 webpages containing CEM in 2018 alone, more than three times the number from 2014 [2]. This is focusing only on websites and does not include other forms of internet-based media-exchange, such as chat-rooms, newsgroups or peer-to-peer (P2P) networks [3] (in particular, it was found that on peer-to-peer network 0.25% of *all* queries were pedophilic [4]).

[1] Corresponding author, E-mail: f.movahedi@gu.ac.ir

As this is an international problem, government agencies such as INTERPOL have created the International Child Sexual Exploitation image database to track and combat this problem. Similar technologies have also come out from private organizations. Google, working with the National Center for Missing and Exploited Children (NCMEC), has adapted its copyright-centric pattern recognition program used on YouTube, to detect child pornography [5]. Microsoft, also working with NCMEC, has created an algorithm called PhotoDNA for detecting modified versions of images, which they have made available for free to law enforcement who deal with child exploitation online and offline [6].

Research has also focused onto the problem of combating child exploitation material (CEM). In one such study, P2P-based child exploitation was examined, for the purposes of developing a filter that can be used to detect queries as pedophilic [7]. In another study, the eDonkey P2P network was studied for the purpose of profile construction based on users' search terms, after which users were classified into those who prefer pedophilia (prepubescent, generally under age 11) vs. those who prefer hebephilia (pubescent, generally age 11 to 14) content [7]. Websites have also been studied through the retrieval and mapping of large networks of websites that host CEM, for the purposes of determining structure [9], key players within the networks [10], and the disruption of those networks through the removal of certain nodes [11].

P2P tends to share files directly and can be queried with keywords and hash values [7] resulting in CEM that is relatively easy to find while web-based repositories require digging and exploration to find such content. Thus mapping P2P can actually be easier while analyzing websites requires exhaustively mapping both the size and content on it. This paper focuses on website-based content.

Although a lot of money and time has been invested into various forms of combating online CEM, the problem is nowhere near under control and there is a dearth of statistics on how this expenditure translates into victims being rescued and offenders getting prosecuted. This is not a comment against law enforcement but rather speaks to the extent of the problem. With so many websites containing child sexual abuse images and videos, and the limited resources available to various organizations to combat the problem, there needs to be continued efforts to automate and simplify the process of selecting and prioritizing targets for the purpose of the criminal investigation. While the cessation of online child exploitation and the distribution of CEM are unlikely, to prioritize, investigations need to take in to account the severity and exposure of the content rather than simply their presence.

One of the studied issues in analyzing the networks is the removal of some important nodes or the hubs, the nodes with the highest connectivity, which separate a complex network into some disconnected components [12-15].

Finding an optimal strategy for disrupting online networks that deal with CEM depends on the specific goals but is a major task regardless. A good attack strategy will cause the largest

disruption of the network by selecting the most important nodes termed key players [11]. In the context of CEM, key players could be measured by a combination of factors, such as the site's influence (possibly measured by the number of other sites linking to them), whether the site is a hub (contains a lot of links to other sites), or the amount of content on it (measured in terms of the number of images or videos it contains). Targeting and removing the sites that score high along these factors would allow law enforcement to make the best use of their limited resources. In that sense, key players represent nodes which are among the goals when it comes to disruption of an online network [11].

In identifying appropriate attack strategies, it is important to consider the topology of the networks. Online networks have two important structural features: Power-law distribution and Small-world properties [10]. The complexity of online networks resides in the small average path lengths among any two nodes (i.e., the small-world property), along with a large degree of local clustering (i.e., the power law distribution). In other words, some special nodes of the structure develop a larger probability to establish connections pointing to other nodes. This introduces problems when finding a node to remove in these networks. Scale-free networks are dominated by a relatively few, highly connected nodes, with the vast majority of nodes being poorly connected [16]. The simplest attack strategy one can consider consists of the random nodes from the network. However, scale-free networks can be regarded as a bounding case of heterogeneous networks, thus, for efficiently attacking a heterogeneous network using a random attack a large number of nodes need to be selected for removal. Therefore scale-free networks are extremely resistant to disruption by random deletion of nodes and targeted attacks must be used on these networks to effectively disrupt them [17].

Methods to disrupt online networks have been proposed in the past. For example, hub attacks, bridge attacks, fragmentation attacks and random attacks have been applied to a large online network of websites hosting child exploitation content [10]. The removal of websites identified by these attack strategies followed a greedy sequential process which i) considered a measure m, ii) identified the website that scored highest according to m, iii) removed the identified website from the network, then iv) re-analyzed the network to identify the next top website according to m. This process was repeated until the top websites were eliminated. The impact of the attack strategies was assessed on four measures of disruption: density, clustering coefficient, average path distance and distance-based cohesion [10].

Identifying the best attack strategy can help improve the efficiency of law enforcement resources when it comes to combating online child exploitation. This paper proposes a new Principal Component Analysis (PCA) -based method to improve the process by which nodes are identified based on their importance and influence within an online network of child exploitation websites. PCA is a useful statistical technique with a common technique for finding patterns in data of high dimensions. It has multiple applications in chemistry, biology, epidemiology, finance, Medical [18]. For example, in [18], the PCA method is investigated to find the most relevant topological and disease parameters in an epidemiologic model. One of the important problems in

the analysis of complex networks is to find the key nodes in the network. In [20], an approach based on non-linear principal component analysis is proposed to identify the top important nodes. Zhang et al., proposed a statistical method based on the PCA algorithm to evaluate the importance of a node in complex networks [21]. Some studies proposed algorithms based on the PCA method for identifying the important nodes in complex networks in the different fields can be found in [21-25].

In this paper, first, network data is collected from the Internet using a custom-written web-crawler (Section 2.1) after which PCA (Section 2.2) is used to identify key players (Section 2.3) which are then removed (Section 2.4). Results indicate that the proposed PCA method outperforms existing network-attack strategies (Section 3) which would have important implications for law enforcement (Section 4).

2. Methods

To introduce the novel PCA method to disrupt child exploitation networks, first, a sub-network centered on online child exploitation material is extracted from the Web using a custom-written web-crawler called CENE (Section 2.1). Using an adjacency matrix and a Laplacian matrix representation of the network, Principal Component Analysis (Section 2.2) is used to identify key players within the network (Section 2.3). Finally, the proposed method is used to formulate an attack strategy (Section 2.4).

2.1. Web-Crawler

Information on the scale and scope of online child exploitation material can be discovered by studying the websites that contain this type of content. One strategy for doing this is to manually visit the websites under study, read the webpages, establish the content of each webpage, then search for links leading to other webpages, and finally map out the hyper-links between those pages. Analyzing the website manually might lead to accurate conclusions about the content, but studying a large website with thousands of pages in this fashion is infeasible. Similarly infeasible is the manual creation of a map of the inter-linkages between the pages for the purposes of social network analysis. Due to the large scale of the problem, the data collection, and analysis, must be performed by computers.

Web-crawlers are the tools used by all search-engines to automatically navigate the Internet and collect information about each website and webpage. They, given a starting webpage will recursively follow the links out of that webpage, until some user-specified termination conditions apply. During this process, the web-crawler will keep track of all the links between other websites and (optionally) eventually follow them and retrieve those as well. There is much standalone web-crawler software available on the Internet, such as 'Win Web Crawler[2],

[2] http://*www.winwebcrawler.com*

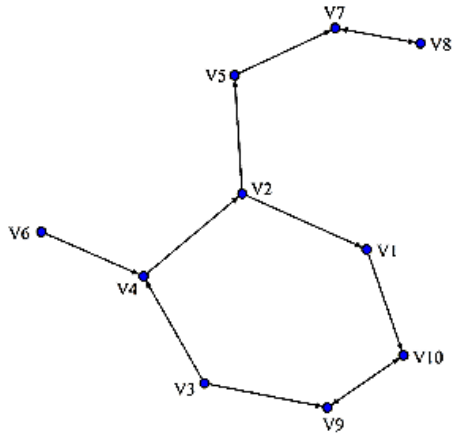
WebSPHINX[3] or Black Widow[4], and some could be used to capture the content of a website onto the machine of the investigator for evidentiary purposes.

However, there are several problems with off-the-shelf web-crawlers. First, most such web-crawlers will save all the content onto the hard-drive, which might work for law-enforcement, who are allowed to do this, but for researchers studying this problem, it is against Canadian law to store such content on a hard-drive even temporarily. All analysis must be done in-memory. Second, when looking for targeted content, off-the-shelf web-crawlers do not perform an adequate job as the search process must be guided by conditions, such as the presence of a child exploitation image on a website, or multiple child exploitation-specific keywords within the body of the webpage. The presence of a child exploitation image can be detected using hash values such as MD5, SHA1 or PhotoDNA, which translate an image into a series of numbers (a hash-value) that are then be compared to a database of previously identified child exploitation images.

To bypass these problems, a custom-written crawler, called Child Exploitation Network Extractor (CENE) was used to collect information on these online child exploitation networks. As CENE visits each page, it captures the contents of the webpage for later analysis, while simultaneously collecting information about the webpage and making decisions on whether it contains child exploitation material, or not. All processing such as hash value and keyword frequency calculations are done in-memory, with the resulting information stored in a central database. For each network extracted, features are collected about the contents of each webpage, and the links between them. This information is stored at the *webpage* level, and then aggregated up to the *website* level. For example, all pages on www.website.com are visited, analyzed, and statistics calculated for each page. After this process is done for all webpages of interest, then all statistics are aggregated up to a single set of statistics for the website www.website.com itself [9].

The data collection was started with seed points collected through popular search engines using previously identified CEM-specific keywords. The URLs presented as the search results were then inputted into CENE, for it to recursively follow links out of the seeds and scan the entire website. These websites were not visited by humans before being analyzed by CENE, thus the size and structure of the webpage were not a factor in whether that website was used as a seed, or not. For each page, CENE decided whether the page should be considered as CEM, and if so, all links on the page were added to a queue for later retrieval. The criterion for this decision was set at 1) the presence of one known child exploitation hash value (i.e., an image already encountered by law-enforcement and known to be an illegal child exploitation image) and/or 2) the presence of seven child exploitation keywords (specified below).

[3]http://www.cs.cmu.edu/~rcm/websphinx
[4] http://sbl.net



a) A sample network of 10 nodes.

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

b) The corresponding adjacency matrix.

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ -1 & 2 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & -1 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

c) The corresponding Laplacian matrix.

Figure 1: A sample network and its corresponding adjacency matrix, and Laplacian matrix.

The web-crawler continues to examine websites until it does not find any more such sites, meaning, none of the sites analyzed link to any new sites containing CEM. A criterion used in the crawling process was the exclusion of websites known to not contain child exploitation content. These websites were based on a list of the most popular websites (e.g., Google) and a list of websites collected during previous data collections that were verified to not contain child exploitation content. The resulting network contains information about the number of images (overall and known child exploitation), videos, keywords, and linkages.

2.1. Principal Component Analysis

The main objective of this study is the disruption of an online child exploitation network by removing appropriate websites from the network. To do this, two methods are introduced to select some nodes from the network based on Principal Component Analysis (PCA). PCA is a common statistical technique for finding information in data with high dimensions [26]. Advanced topics and technological methods in PCA are given in the book by Jolliffe [27].

PCA compresses data without much loss of information [26, Chapter 2], and mathematically defines a new coordinate system by determining the eigenvectors and eigenvalues of a matrix that optimally describe the variance in a single dataset. Correlation between variables in the

dataset corresponds to the degree of variance such that the greatest variance by any projection of the data comes to lie on the first coordinate which is called the first principal component [27, Chapter 3]. PCA involves a calculation of a covariance matrix of a dataset to minimize the redundancy and maximize the variance [26, Chapter 1], in the process of finding a linear mapping of a dataset to a dataset of lower dimensionality.

In computational terms, the principal components are found by calculating the eigenvectors and eigenvalues of the data covariance matrix. This process is equivalent to finding the axis system in which the covariance matrix is diagonal. Matlab software is used for computing of PCA algorithm.

In this study, the input data set of PCA is an $n \times n$ matrix M which represents the structure of the network. Then the mean of the data is subtracted from each data value in order to obtain a data set with zero means. The resulting matrix is named $M - \bar{\mu}$. The covariance matrix of $M - \bar{\mu}$ is then calculated, and is given by:

$$C^{n\times n} = \left(c_{ij}\,, c_{ij} = cov\left(X_i, X_j\right)\right),$$

where $C^{n\times n}$ is a matrix with n rows and n columns, X_i is the i^{th} dimension and $cov\left(X_i, X_j\right) = \frac{\sum_{k=1}^{n}(X_{ik}-\overline{X_i})(X_{kj}-\overline{X_j})}{(N-1)}$, that is the covariance between i^{th} and j^{th} dimensions.

The covariance is a measure to find out how much the dimensions differ from the mean with respect to each other. Then the eigenvalues and eigenvectors of the covariance matrix are calculated. These eigenvalues and eigenvectors show the useful and important information of the data. Noticeably, the eigenvectors are perpendicular to each other. It should be noted that eigenvalues have quite different values. In general, the eigenvectors from the covariance matrix are ordered by eigenvalue, highest to lowest. This gives the components in order of significance. At this time, one could decide to ignore the components of lesser significance. By taking the eigenvectors of the covariance matrix one can extract vectors that characterize the data. The final step in PCA is to choose the first principal component (eigenvector) and to form a feature vector.

The eigenvector with the largest eigenvalue is the direction of greatest variation with the maximal variance of the data set [26] and the second largest eigenvalue is the (orthogonal) direction with the next highest variation, and so on [26]. So, in this study, only the eigenvector corresponding to the highest eigenvalue is considered a feature vector. Let $\propto_1$be this eigenvector with the norm of 1. Define $B_{1\times n} = \propto_1^T \times (M - \mu)^T$. The vector B approximately consists of all information pertaining to principal data [26].

2.2. PCA Metrics for Identifying Key Players

In mathematical modeling of complex networks, the representation of the networks can denote by the Adjacency matrix and Laplacian matrix of the network. For analyzing and studying the networks, it is sufficient to compute some measures in these networks. According to the definition of the Adjacency matrix and its properties, some measures are usually used for analyzing the network such as Clustering coefficient, Average path length and Degree distribution. And in the network modeling with the Laplacian matrix, the measures such as centrality and connected components are used.
We introduce two strategies using PCA, named PCA_A and PCA_L, to identify the most important key players in an online exploitation network. For this purpose, the adjacency matrix and the Laplacian matrix of a network are considered input data sets to the PCA algorithm.

The adjacency matrix, named A, of a network G consisting of n nodes is the $n\times n$ matrix where $A(i,j) = 0$ if nodes i and j are not connected, otherwise $A(i,j) = 1$. The adjacency matrix is represented to show nodes and connectivity between them.

The Laplacian matrix of a directed network $L = (l_{i,j})_{n\times n}$ is defined as $L = D - A$ where $D = diag(d_1, d_2, \dots, d_n)$ is the matrix, the diagonal matrix which in d_i denotes the out-degrees or in-degrees of the nodes in the network [27]. From the definition it follows that:

$$l_{i,j} = \begin{cases} d_i & if\ i = j \\ -1 & \text{if i} \neq \text{j and } v_i \text{is adjacent to } v_j \\ 0 & otherwise \end{cases} \quad (2.1)$$

The Laplacian matrix represents the nodes and connectivity between the nodes of the network and shows the network structure. It should be noted at once that loops have no influence on L [28, 29]. The Laplacian matrix of a graph and its eigenvalues can be used in several areas of mathematical research and have a physical interpretation in various physical and chemical theories. The Laplacian spectrum is more important than the adjacency matrix spectrum [28].

Two methods are proposed above to distinguish the key players in a network. In the PCA_A strategy, the adjacency matrix (A) of the network is considered as the input into the PCA algorithm. In the PCA_L strategy, the Laplacian matrix of the network is considered as M. Each column of M yields the information relevant to the connectivity of each website in the network. So, the one node of the network corresponding to the maximum entry of B can be considered a key player.

To clarify the issue, we provide a simple example. We consider a sub-network with 10 nodes from the initial extracted network by CENE (Figure 1(a)). Assume the nodes of the network are labeled with v_i with $1 \leq i \leq 10$ and consider vector $V = (v_1, v_2, \dots, v_{10})$ as the nodes' vector.

The adjacency and Laplacian matrices of this network are denoted by A and L where the Laplacian matrix is calculated by equation (2.1) (see Figure 1(b and c), respectively).

For the PCA_A strategy (see Section 2.2), the matrix $(A-\mu)$ is calculated in which μ is the mean of the data in each column. Let α_1 be the eigenvector associated with the largest eigenvalue from the covariance matrix of $(A-\mu)$. So, the vector $B_{1\times 10}$, the output of the PCA algorithm, is as follows

$$B = [0.382 \quad 0.383 \quad -1.1601 \quad 0.1559 \quad 0.381 \quad -0.531 \quad 0.156 \quad 0.382 \quad 0.382 \quad -0.521].$$

The maximum entry of this vector is 0.383 which corresponds to the 2nd entry of the nodes' vector. So, we select the node v_2 as a key player. According to matrix A, it is also worth mentioning that node v_2 has the maximum out-degrees in this network (based on the hub strategy). It means that this website may provide abundant access to materials in the network. So, it is an appropriate node in the network that removing it can help for disrupting this network. For the PCA_L strategy, the Laplacian matrix (L) of the network is considered as the input data set M into PCA, resulting in vector $B_{1\times 10}$

$$B = [-0.736 \quad 1.892 \quad 0.949 \quad -1.116 \quad -0.617 \quad 0.192 \quad 0.043 \quad -0.326 \quad -0.645 \quad 0.363].$$

The maximum entry of vector B is 1.892 which corresponds to the 2nd entry of the nodes' vector. So, the PCA Algorithm selects the node v_2 from the network.

2.3. Key Player Removal

Key players can be defined as nodes with large volumes of content. However, we focus on the phase before this, when the offender is seeking out this content and is exploring the network of CEM websites, hopping from one website to another seeking new material. As a result, for the purposes of this paper, we define a key player as a website that is important not in terms of content, but in terms of *network* position. Thus, we consider some measures for the evaluation of the network. Since both the Adjacency and Laplacian Matrices are used to represent some properties of networks thus, we consider them for investigating the resulting network in any steps.

In each iteration, one node is identified by PCA as being the most important in the network. The attack scenario presented in this paper is greedy, a node is removed from the network after which the remaining network is reanalyzed to identify the next best node. This process is repeated until removing key players by the PCA approach, the network is disrupted. Since our aim is to disrupt the initial network while removing the minimum number of nodes, our algorithm selects one node at any step. Finally, the resulting network is examined, and the obtained results are compared to the outcome measures of other attack strategies.

For comparing the results of the presented strategies to other strategies, we consider various attack strategies. These attack strategies involve hub attacks (using the measure of degree

centrality), bridge attacks (using the measure of Betweenness), network capital [11] (where the node that contributes the most content is selected), and random attacks (where each node has an equal chance of being targeted).

In networks, centrality is the measure of how important a node is in the network. Betweenness centrality is a measure of centrality in the network. In the bridge attack method, an important node based on Betweenness centrality is selected. Therefore, we use the PCA_L method to compare the proposed approach based on the PCA method by considering the Laplacian matrix as the input of the algorithm to study strategies, especially the bridge attack.

Continuing the example from Section 2.3, after removing the node v_2 from the network, PCA_A identified node v_9 as the next key player. PCA_L identified nodes$\{v_2, v_3\}$. We consider other attack strategies for this network to identify two nodes in each step. For this purpose, these strategies include Hub Attack, Bridge Attack and Random Attack. The original and resulting networks using different attack strategies are shown in Figure 2. It can be seen from Figure 2(b) and Figure 2(f) that the network became fragmented into three separate components following both the PCA_A and PCA_L attack strategies. So, in each of the components, it would be harder to discover the other small connected components, as no link leads from one connected component to another. While the graphs of Hub Attack and Bridge Attack have one component with two isolated nodes (see Figures 2(b and c)). The set of selected nodes by Hub Attack and Bridge Attack from initial network are $\{v_4, v_7\}$ and $\{v_2, v_7\}$, respectively. Also, as seen in Figure 2(d) the network has only one component after removing two nodes v_8 and v_9 by Random Attack.

The impact of removing these websites by the proposed strategy is then examined on several outcome measures: density, clustering coefficient, average path length, diameter and the number of connected components in the resulted network. Density is calculated by dividing the number of existing links in the network by the maximal number of links [29]. So, the changes in density correspond to the changes in the number of links. The other outcome measure included is the network clustering coefficient which is the average density of the neighborhoods of the websites in a network [29]. In other words, it is defined as the probability that two randomly selected neighbors are connected to each other. The average path length, defined as the average number of links along the shortest paths between two nodes [29], is examined for all pairs of nodes in the network.

In this paper, we use Matlab for computing and coding the attack simulation. For obtaining the average path length of the resulting networks, we compute the shortest distance between pairs of nodes in the network using Matlab's *graphallshortestpaths*(*G*) function, based on Johnson's algorithm [31]. In this study, for two nodes v_i and v_j that have no other connections, we considered the shortest distance $d(v_i, v_j) = 0$. So, if the size of the network is n then the average path length (avg) of the network is obtained as follows

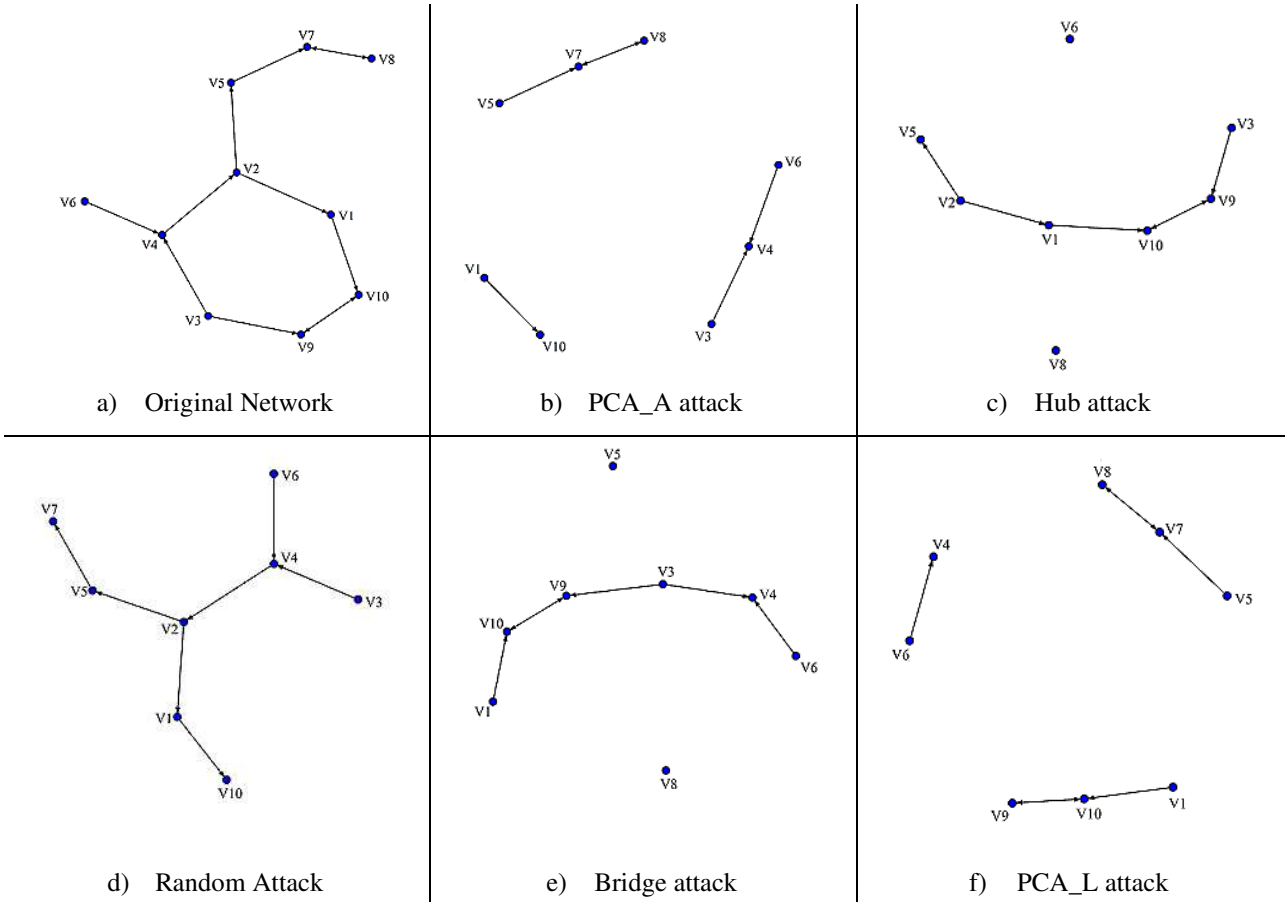
$$avg = \frac{\sum_{i \neq j} d(v_i, v_j)}{n(n-1)}.$$

The maximum distance between any two vertices in a network is called the diameter of the network. The impact on network connectivity of selecting and removing targeted nodes as measured by changes to network diameter is evaluated in [32].

For instance, if the nodes which are most highly connected are removed through the Hub attack then, the network diameter increases rapidly and the resulted network fragments into smaller components or subgraphs [32].

3. Results and Discussion

The main goal of this study is to determine a method to find the websites that should be prioritized by law enforcement agencies involved in combating child exploitation. In this study, an online network is extracted using CENE, a web-crawler tailored to follow the links out of and into child exploitation websites when given a specific set of starting websites [11]. CENE started to explore a network from a single child exploitation website that was found through extensive searches on Google. CENE would not follow links out of a single webpage if that webpage did



a) Original Network b) PCA_A attack c) Hub attack

d) Random Attack e) Bridge attack f) PCA_L attack

Figure 2: Resulting network after various attacks.

not meet certain criteria associated which child exploitation (at least one known image, or at least seven pre-identified keywords related to child exploitation). These were words that were provided by the RCMP as well as those used in previous research [3]. These words are included $qwerty$, $qqaazz$, $ptsc$ and $pthc$. This category of keywords consisted of twenty-seven words, which, to the best knowledge of the authors, were still valid during data collection.

With these starting parameters, CENE identified a network of connected websites that contained at least one page matching the requirements and stopped expanding the network when none of the webpages linked to child exploitation material anymore (i.e., the links led to webpages off topic). This resulting network contained 177 nodes and 915 edges, where nodes are defined to be entire websites (web-domains), and the directed edges represent the links pointing from one node (website) to another. A node in the network is labeled with i where $1 \leq i \leq 177$. The initial network density, clustering coefficient, average path length and network diameter are 0.0294, 0.5095, 0.0030 and 6 respectively.

First, matrix A (the adjacency matrix) and matrix L (the Laplacian matrix) are obtained from the network. The PCA algorithm selects a node from the network using two matrices A and L in each step. After removing the selected node, the algorithm calculates outcome measures of the resulting network and updates the new network to identify the next appropriate node. This process was repeated until the network breaks up into components for both the strategies.

In the PCA_A strategy, the adjacency matrix A, which in this context contains the number of links one website has pointing to another, was the input into the PCA algorithm. Results (see Table 1) show the following changes after removing five selected nodes such that the first disruption happens. Density fell by 26.5% from 0.0294 to 0.0216 while the number of links dropped from 915 to 687 indicating that by removing only five (2.8%) nodes, 228 (24.9%) of the links were removed. Although this was expected, as this attack focuses on the removal of the nodes with the most links, the amount of damage caused was more severe than expected. The average path length decreased by 26.7% from 0.0030 to 0.0022, meaning that with removing the links between some websites, there is no path between some nodes in the network. Therefore, according to the definition of avg, this value is reduced. Finally, the clustering coefficient in this network decreased from 0.5095 to 0.4520 (-11%), each node was now *less* embedded in the network than before. Thus, through the removal of just five nodes we were able to break 25% of the active links between them, resulting in a much smaller and partially disconnected graph (see Figure 3(a) vs. 3(b)) making it much more difficult for individuals to reach other websites.

As a second experiment, with the PCA_L strategy, the Laplacian matrix L is given as the input into the PCA algorithm. Since websites with many in-degree ties may be considered more important, a website can easily link to others, but it may not be relevant or interesting enough to receive links from other websites, thus in-degree for any of the nodes in the Laplacian matrix are considered. After selecting and removing five nodes the following changes to the network structure were observed. First, density fell by 25.8% from 0.0294 to 0.0218, indicating a similar

amount of damage to the network as PCA_A. The clustering coefficient decreased from 0.5095 to 0.4588 (-10%), and each node was now *less* embedded in the network than before, even when compared to PCA_A. Finally, the average path length fell by 16.8% from 0.0030 to 0.0025 (see Table 2). Overall, PCA_L seems to have severed almost the same number of links (228 in PCA_A vs. 236 in PCA_L) but it did so in such a way that the average path length dropped by only 16.8% (as opposed to 26.7% in PCA_A) meaning that, in the context of a network of child exploitation websites, the average user moving from one website to another would need to click through more websites to reach the destination site while having fewer links available to them to do so. Thus PCA_A significantly increased the difficulty of finding websites within the network, even compared to PCA_L.

Indeed, the aim of this study is to find the strategy which selects the minimum nodes (websites) and will cause the largest disruption to the network. The removal of websites identified by the proposed attack strategies followed a sequential process in which one node is identified by the PCA algorithm as being the most important in the network. Then, a node is removed from the network after which the remaining network is reanalyzed to identify the next best node (i.e., a greedy approach). Table 1 and Table 2 show the changes in the resulting network after removing the node where the PCA algorithm is selected in each step (n). Also, n is the number of selected nodes by PCA and it is the number of removed nodes in the resulting network. According to these tables, for $n = 0, 1, 2, 3, 4$ the variations were reduced in the network after each step, but the removal of the nodes with the most links occurs in step $n = 5$ disruptions of this network. According to Figure 3, the active links between these nodes (websites) were removed and resulting in a smaller disconnected graph.

Table 3 shows the results of calculating the outcome measures after removing five nodes using any of the other existing strategies (Hub or Bridge attacks, for example). Results show that for different outcome measures, some attack strategies are less effective in disrupting the network. Furthermore, the effectiveness of these various attacks varied with the different goals, such as reducing density, clustering and reachability. For example, if the goal is to delete as many links in a network as possible (i.e., reduce density), then the Attack Bridge is the most effective strategy since it led to a reduction of 18% in the density of our test network. However, using the strategy proposed in this paper the density decreased by 26.5% and 25.8% using PCA_A and PCA_L, respectively. It means that the proposed strategies are even more effective in decreasing the number of links.

To reduce a node's embeddedness in a tight-knit component of the network (clustering), the results (Table 3) show that the current attack strategies except hub attack increase the value of this measure while PCA_A and PCA_L decrease it by 11% and 10%, respectively. Also, this measure is decreased in the hub attack a similar amount to PCA strategies.

Table1: The results of outcome measures of obtained network after removing one node by PCA_A strategy in each step

Steps of running PCA	Density of network	Clustering coefficient	Average path length
Step 0	0.0294	0.5095	0.0030
Step 1	0.0276	0.4990	0.0023
Step 2	0.0260	0.2817	0.0023
Step 3	0.0244	0.4612	0.0023
Step 4	0.0229	0.4608	0.0023
Step 5	0.0216 (↓ 26.5%)	0.4520 (↓ 9%)	0.0022 (↓ 26.7%)

Table 2: The results of outcome measures of obtained network after removing one node by PCA_L strategy in each step

Steps of running PCA	Density of network	Clustering coefficient	Average path length
Step 0	0.0294	0.5095	0.0030
Step 1	0.0277	0.4945	0.0030
Step 2	0.0261	0.4813	0.0029
Step 3	0.0246	0.4749	0.0028
Step 4	0.0232	0.4669	0.0026
Step 5	0.0218(↓ 25.8%)	0.4588 (↓ 10%)	0.0025 (↓ 16.8%)

Table 3: The results of outcome measures of obtained network after removing five nodes by various attack strategies

Attack strategies	Density of network	Clustering coefficient	Average path length	Diameter	Changes of the number of CC
Hub Attack	0.0285 (↓ 3%)	0.4520 (↓ 11%)	0.0021 (↓ 30%)	7 (↑ 16%)	2 ↑
Bridge Attack	0.0240 (↓ 18%)	0.5347 (↑ 4.9%)	0.0016 (↓ 46.6%)	7 (↑ 16%)	4 ↑
Network Capital method	0.0305 (↑ 3%)	0.5122 (↑ 0.52%)	0.0032 (↑ 6.6%)	6 (No-change)	4 ↓
Random Attack	0.0299 (↑ 1%)	0.5158 (↑ 1.2%)	0.0034 (↑ 13.3%)	6 (No-change)	5 ↓
PCA_A	0.0216 (↓ 26.5%)	0.4520 (↓ 11%)	0.0022 (↓ 26.7%)	7 (↑ 16%)	2 ↑
PCA_L	0.0218 (↓ 25.8%)	0.4588 (↓ 10%)	0.0025 (↓ 16.8%)	7 (↑ 16%)	2 ↑

Results show that the bridge attack decreased the average path length the most, and thus was the most successful attack among all of the attack strategies, even better than PCA_A or PCA_L (note however that in the other measures the bridge attack was significantly inferior to both PCA_A and PCA_L). On the other hand, this method increases the clustering coefficient measure by 4.9%. So, overall, it can't be the most effective strategy for disrupting this network.

There are some isolated nodes in the network after running the proposed attacks (PCA_A and PCA_L). Since the shortest path length for any isolated node with others is equal to zero in our algorithm so, one can expect to decrease this measure using both PCA strategies. However, the PCA_L method has less reduction than the PCA_A method.

We measure the network diameter and the changes in the number of connected components (CC) to ensure the integrity above discussion.

As for the network diameter, when the targeted nodes are removed, the diameter of the network increases and the network breaks into isolated connected components. This occurs because when deleting these nodes, the heart of the network disturbs, whereas the random attack is most likely not. The results are shown that all of the attack strategies increase the network diameter after removing five targeted nodes by 16% except two for the random and network capital strategies. According to Table 3, after the removal targeted nodes in any strategy, 4 connected components are added to the resulted network by Bridge attack. By PCA methods and Hub attack, 2 connected components add to the network.

In this study, decreasing the average path length is significant if the diameter and the changes in connected components increase. Because after removal nodes, the resulted network contains some connected components of a small size.

The results are shown that the number of connected components in the resulted network after two random and network capital attacks decreased. According to the obtained results of other measures for these two strategies, it is clear to see that the reduction of the number of connected components is due to selecting isolated nodes.

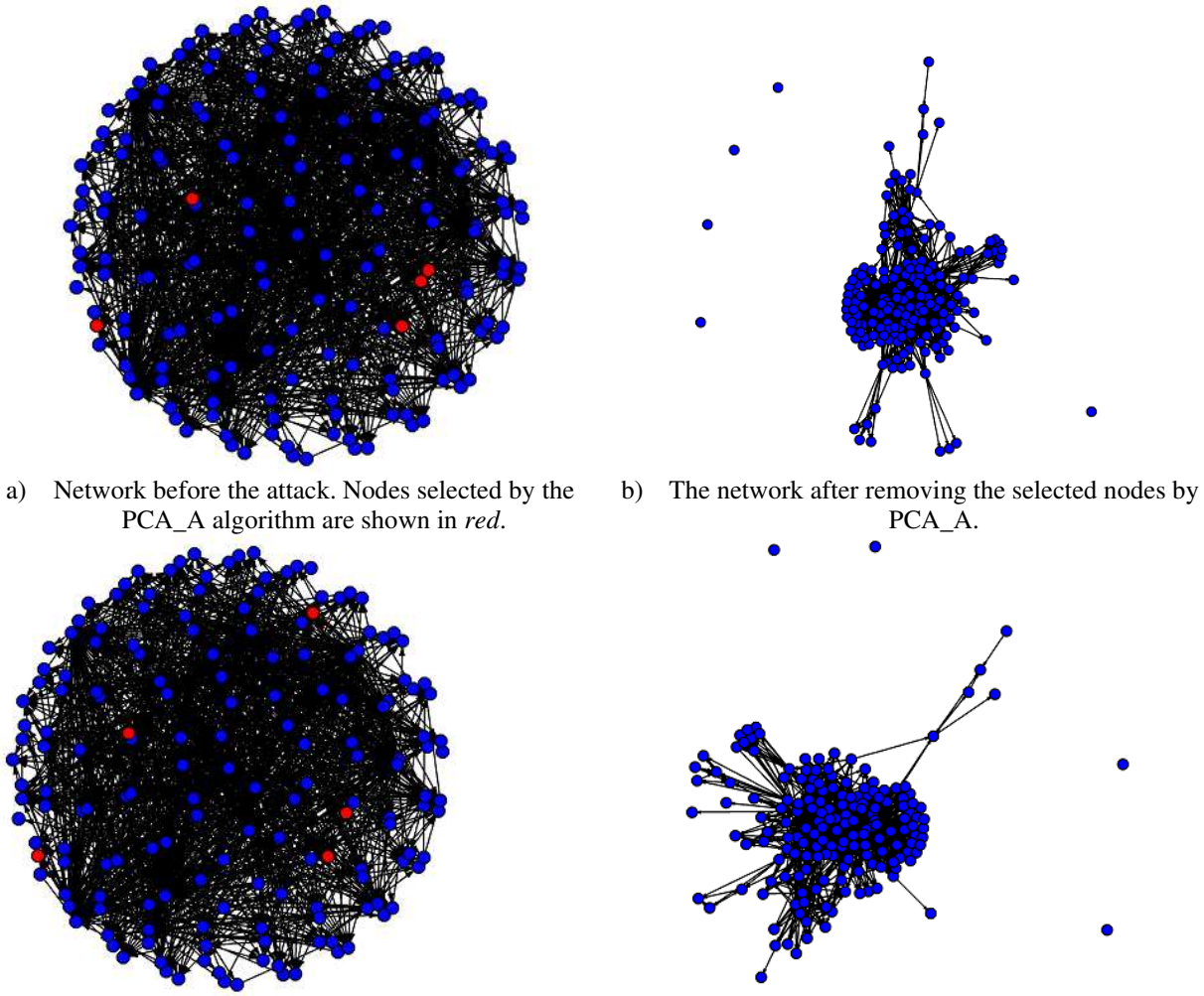
Figure 3 shows the before and after the process by which the original network is changed when the websites are removed by PCA_A and PCA_L. Most of the targeted websites are located inside the original network where they have the most influence on the transmission of information to other websites (see the red nodes in Figure 3(a and c). In addition, as seen in Figure 3(b), the network is now fragmented into one component with five isolates while density decreased by removing 228 edges from the initial network. Also, in Figure 3(d), the resulting network has one component and four isolated nodes after selecting and removing five nodes by the PCA_L strategy. These five nodes also are located inside the initial network. It is also worth mentioning that there are four common nodes between those selected by PCA_A and PCA_L.

Table 4: The results of outcome measures of obtained network after removing 20 nodes by various attack strategies

Attack strategies	**Density of network**	**Clustering coefficient**	**Average path length**	**Diameter**	**Changes of the number of CC**
Hub Attack	0.0082 (↓ 72%)	0.2456 (↓ 51%)	0.0006 (↓ 80%)	5 (↓ 16%)	8 ↑
Bridge Attack	0.0144 (↓ 51%)	0.5251 (↑ 1%)	0.0004 (↓ 86%)	7 (↑ 16%)	9 ↑
Network Capital method	0.0178 (↑ 39%)	0.5123 (↑ 0.5%)	0.0019 (↑ 36%)	6 (No-change)	7 ↓
Random Attack	0.0319 (↑ 8%)	0.5049 (↑ 9%)	0.0024 (↑ 20%)	6 (No-change)	8 ↓
PCA_A	0.0153 (↓ 47%)	0.2396 (↓ 53%)	0.0008 (↓ 73%)	7 (↑ 16%)	9 ↑
PCA_L	0.0080 (↓ 73%)	0.2429 (↓ 52%)	0.0005 (↓ 83%)	7(↑ 16%)	9 ↑

It is clear that one may select more nodes and repeat the proposed algorithms for selecting and analyzing the obtained networks in steps $n > 5$ for this dataset. Table 4 shows the results after removing 20 nodes by various attack strategies.

The results are shown both strategies (PCA_A and PCA_L) have better performance than the four proposed strategies after selecting and removing more nodes. According to Table 4, the PCA_L strategy yielded significant improvements over the existing methods versus others. Although the hub attack did yield results close to both PCA strategies in some measures, one strategy is more successful than existing strategies for disrupting the network that significantly changes all outcomes.
According to the obtained results, we can use the PCA_L strategy for disrupting this dataset (see Table 4). This strategy is based on the Laplacian matrix that representation of the network and its properties.



a) Network before the attack. Nodes selected by the PCA_A algorithm are shown in *red*.

b) The network after removing the selected nodes by PCA_A.

c) Network before the attack. Nodes selected by the PCA_L algorithm are shown in *red*.

d) Resulted network after removing five nodes using PCA_L.

Figure 3: Network before and after attack by the PCA_A and PCA_L strategies.

4. Conclusion

This experiment attempted to identify the most important nodes in an online network containing child exploitation images, extracted from the Internet. The network was extracted using CENE, a custom-written web-crawler that can identify known child exploitation images through their hash values and focuses on them for the purposes of network extraction. Agencies that combat this problem have limited resources, thus for the purposes of time savings and cost reduction, it is important to select the most efficient attack strategies. There are multiple existing strategies for disrupting online child exploitation networks that are dependent on some properties of networks, such as a hub attack strategy which depends on using the measure of degree centrality. The aim of these studies is to select the strategy that will cause the largest disruption to the networks themselves.

In this paper, six attack strategies were used to attack the structure of a single online network of child exploitation websites. The goal was to determine whether the two proposed Principal Component Analysis techniques presented an advantage over four previously established attack methods (hub, bridge, network capital and random). Considering the adjacency matrix (PCA_A) and Laplacian matrix (PCA_L) in the PCA algorithm, each column of these matrices yields the information on connectivity relevant to each website in the network. The nodes which were selected by PCA correspond to websites yielding significant information to the network. The results show both the two strategies (PCA_A and PCA_L) have better performance than the existing four strategies when it comes to disrupting the network, although the Hub attack yielded the results close to both PCA strategies. Overall, however, results shown after selecting and removing more nodes of the network using various strategies, the PCA_L strategy yielded significant improvements over the existing methods.

Although the data collection aimed at finding as many of these websites as possible, there are two limitations to the data collection, which might impact the generalizability of this study. First, the data collection was started with seed points which were identified through keyword searches through various search engines, and data collection proceeded from one website to another through the linkages available within the websites. While websites containing similar material do tend to link together, any website which was not linked would have been missed using this strategy. This limitation was mitigated by having multiple seed starting points, but regardless, if a website is not linked to, then it was not visited and thus not included in the data collection. Second, this research is focused on open internet networks only, thus no other types of networks were included, such as dark-web sites, file-sharing networks, or private password-protected websites.

Future work should look at the effectiveness of these attacks in other types of networks, not just different networks extracted from the Internet, but also networks that do not share the Internet's Power-law distribution and Small-world properties. The resiliency of nodes of various importances could also be investigated with the help of law enforcement. If law enforcement

were able to incorporate such attack strategies into their selection methodologies, and then actually remove them from the Internet, CENE could be used to monitor the resulting network for a possible redistribution of the content and/or importance. Do more important nodes get more prominent, or do the smaller websites take up the opportunity and become more prominent within the network structure? Or perhaps, like a hydra, many other websites are created to fill the void?

5. Declarations

Ethical Approval and Consent to participate Not applicable.

Human and Animal Ethics This work does not contain any studies with human participants or animals.

Consent for publication Not applicable.

Availability of supporting data Not applicable.

Conflict of Interests The authors declare that they have no conflict of interest.

Funding Not applicable.

Author contributions Fateme Movahedi introduced the main approaches in the proposed model. Richard Frank provided and collected the network data. Fateme Movahedi prepared tables of the numerical results and figures. Fateme Movahedi and Richard Frank investigated and wrote the section of results and discussions. All the authors read and approved the final manuscript.

Acknowledgments Fateme Movahedi would like to thank Dr. Vahid Dabbaghian and Dr. Piper Jackson for their constructive suggestions and helps.

Authors' information

Fateme Movahedi received the MS degree in Mathematics, from the University of Tehran, Iran, in 2009, and the Ph.D. degree in Mathematics from Shahid Beheshti University, Tehran, Iran, in 2013. She is an assistant professor in mathematics, at Golestan University. Her research interests include Social Networks, Mathematical modeling, and applicational combinatorics.

Richard Frank is Associate Professor in the School of Criminology at Simon Fraser University (SFU), Canada and Director of the International CyberCrime Research Centre (ICCRC). Richard completed a PhD in Computing Science (2010) and another PhD in Criminology (2013) at SFU. His main research interest is Cybercrime. Specifically, he's interested in researching hackers and security issues, the dark web, online terrorism and warfare, eLaundering and cryptocurrencies, and online child exploitation. He is the creator of The Dark Crawler, a tool for collecting and analyzing data from the open Internet, dark web, and online discussion forums.

References

[1] E. Engeler (2009) UN expert: Child porn on internet increases.The Associated Press. Retrieved from: http://abcnews.go.com/Technology/wireStory?id=8591118.

[2] Internet Watch Foundation 2018, Report. Retrieved from https://www.iwf.org.uk/sites/default/files/reports/2019-04/ Once%20upon%20a%20 year %20-%20IWF%20Annual%20Report%202018.pdf

[3] M. Latapy, C. Magnien,R. Fournier (2009) Technical report on the Quantification of paedophile activity in a large p2p system. Measurement and Analysis of P2P Activity Against Paedophile Content Project. Retrieved from: http://antipaedo.lip6.fr.

[4] M. Latapy, C. Magnien, R. Fournier, Quantifying paedophile activity in a large p2p system. Information Processing & Management , 49 (2013), 248-263.

[5] M. Shiels (2008). Google tackles child pornography. *BBC News.* Retrieved from: http://news.bbc.co.uk/2/hi/7347476.stm.

[6] Microsoft. (2009). New technology fights child porn by tracking its "PhotoDNA". Retrieved from: https://www.microsoft.com/presspass/features/2009/dec09/12-15photodna.mspx.

[7] M. Latapy, C. Magnien, R. Fournier, Report on Automatic Detection of Paedophile Queries,Measurement and Analysis of P2P Activity Against Paedophile Content project http://antipaedo.lip6.fr, 2006.

[8] S. Hammond, E. Quayle, J. Kirakowski, E. O'Halloran, F. Wynne, An Examination of Problematic Paraphilic use of Peer to Peer Facilities, International Conference on Advances in the Analysis of Online Paedophile Activity, 2009.

[9] R. Frank, B. G. Westlake, M. Bouchard, The structure and content of online child exploitation.Proceedings of the 16th ACM SIGKDD Workshop on Intelligence and Security Informatics (ISI-KDD 2010).

[10] K. Joffres, M. Bouchard, R. Frank, B. Westlake, Strategies to disrupt online child pornography networks. Paper presented at the European Intelligence and Security Informatics Conference 2011, Athens, Greece, 2011.

[11] B. G. Westlake, M. Bouchard, R. Frank, Finding the key players in online child exploitation networks. Policy and Internet, 3(2) (2011) Article 6. doi:10.2202/1944-2866.1126.

[12] I. A. Kovacs, A. L. Barabási, Network science: Destruction perfected. Nature, 524 (2015), 38-39.

[13] R. Albert , H. Jeong, A. L. Barabasi, Error and attack tolerance of complex networks, Nature. 406(6794), (2000) 378-382.

[14] F. Morone, B. Min, L. Bo, R. Mari, H. A. Makse, Collective influence algorithm to find influencers via optimal percolation in massively large social media. Sci Rep (2016) 6:30062.

[15] B. Amiri, M. Fathian, E. Asaadi, Influence maximization in complex social networks based on community structure, Journal of Industrial and Systems Engineering, 13 (3), (2021), 16-40.

[16] A. L. Barabási, The physics of the Web, Physics World, 14 (2001), 33-38.

[17] R. Pastor-Satorras, A.Vespignani, Immunization of complex networks: Physical Review E, 65(2002), 036104.

[18] Y. Mori, M. Kuroda, N. Makino, Nonlinear Principal Component Analysis and Its Applications, Springer, 2016.

[19] P. H. T. Schimit, F.H.Pereira, Disease spreading in complex networks: A numerical study with Principal Component Analysis, Expert Systems with Applications, 97 (2018), 41-50.

[20] S. Basu, U. Maulik, Mining important nodes in complex networks using nonlinear PCA, 2017 IEEE Calcutta Conference (CALCON), (2017).

[21] K. Zhang, H. Zhang, Y. dong Wu, F. Bao, Evaluating the importance of nodes in complex networks based on principal component analysis and grey relational analysis, 2011 17th IEEE International Conference on Networks, (2011) 231-235.

[22] Y J. Jin, K. Xu, N. Xiong, Y. Liu , G. Li, Multi-index evaluation algorithm based on principal component analysis for node importance in complex networks. Networks, IET 1(3), (2012) 108–115.

[23] P. Wang, J. Lu , X. Yu, Identification of Important Nodes in Directed Biological Networks: A Network Motif Approach, PLOS ONE, 9 (2014), e106132.

[24] F. Hu, Y. Liu, Multi-index algorithm of identifying important nodes in complex networks based on linear discriminant analysis, Modern Physics Letters B, 29(03), (2015) 1450268.

[25] F. Hu, Y. Liu, J. Jin, "Multi-index Evaluation Algorithm Based on Locally Linear Embedding for the Node Importance in Complex Networks", DCABES, (2014), 138-142.

[26] L. I. Smith, A tutorial on Principal Components Analysis, Maintained by Cornell University, (2002).

[27] I. T. Jollife, Principal component Analysis, Springer Science & Business Media, (2013).

[28] C. Godsil, G. Royle, Algebraic Graph Theory. New York: Springer-Verlag, (2001).

[29] B. Mohar, The Laplacian spectrum of graphs, in: Y. Alavi, G. Chartrand, O. R. Oellermann, A. J. Schwenk (Eds.), Graph Theory, Combinatorics, and Application, 2, (1991), 871-898.

[30] David Knoke and Song Yang, Social Network Analysis. Quantitative Applications in the Social Sciences, SAGE Publications, (2019).

[31] D. B. Johnson, Efficient algorithms for shortest paths in sparse networks, Journal of the ACM, 24 (1), (1977), 1-13, doi:10.1145/321992.321993.

[32] R. Albert, H. Jeong, A. L. Barabάsi, Error and attack tolerance of complex networks, Nature 406 (2000), 378-382.