August 29, 2019

SPAM IDENTIFICATION USING SET-COVER ALGORITHM ON A CHANNEL SIMILARITY GRAPH

Alexandru Moşoi

Andreas Noever

SPAM IDENTIFICATION USING SET-COVER ALGORITHM ON A CHANNEL SIMILARITY GRAPH

Media content platforms are a highly useful tool for distributing media content and information. However, media content platforms can attract large amounts of spam due to the ease of dissemination of content. Spam can reduce the quality of platform content and can also reduce the appeal of the platform to users. Therefore, identifying and removing spam is an important part of managing media content platforms. Unfortunately, there may be such a significant number of spam channels that using manual human review of each channel can become impractical. A channel may be a platform for a user to upload content. Generally, spam is uploaded through thousands of channels that are created and used by the same attacker. Thus, in many cases, a large number of spam channels contain the same content, making human reviews repetitive and wasteful.

Conventionally, spam is identified, at least in part, through human review of channels. Once identified, the spam can be removed from the platform and the spam channel can be suspended. Some current methods of identifying spam include some amount of automation in conjunction with human review. For example, a label propagation algorithm may be applied to a channel similarity graph. A channel similarity graph may be a graph in which the nodes are channels and the edges connecting nodes represent similarity between channels. A label propagation algorithm may propagate known labels to other nodes based on the frequency of the labels of neighboring nodes. For example, if a channel is connected to more nodes (channels) that are labeled as spam than not spam, the channel may be labeled as spam. However, label propagation can be noisy and precision drops when propagated over multiple edges. Therefore, the entire set of channels may not be accurately assessed for spam. In another example, active learning may be used to identify spam channels. Active learning may query channels that have been human reviewed in order to "learn" how to identify spam channels. In another example, if a channel is related to some number "N" of other channels that have already been human reviewed and identified as spam then the channel can be inferred to be spam without individual human review. However, the channels to review are selected at random and therefore reviews can be inefficient, as explained above.

Using the current methods described above, human reviews may not result in as many spam takedowns as they could have if a more selective process were used. None of the current methods for detecting spam channels minimize the number of human reviews while maximizing spam channel identification. Therefore, a way to select which channels to provide for human review to increase or maximize identification of spam channels is needed.

Presented herein are methods to select potential spam channels for human review using a minimum set-cover approximation algorithm applied to a similarity graph of a group of channels. The use of the set-cover algorithm can maximize the number of spam channels identified while minimizing the number of channels required to be reviewed. The set-cover algorithm is used to identify a small subset of channels which are not yet reviewed by a human reviewer that after review will maximize the number of rejected spam channels. In one example, the set-cover algorithm follows a series of steps, iteratively selecting the best channel or channels to review based on the similarity graph.

The set-cover algorithm is an algorithm intended to select a subset of all nodes of a graph such that every node in the graph is connected to at least one of the selected subset of nodes. A *minimum* set-cover algorithm attempts to minimize the number of selected nodes in the subset, so that the smallest possible number of selected nodes can cover the entire set. One minimum set-cover algorithm uses what is known as a greedy solution. A greedy solution is an iterative process in which the best node at that point is selected without looking forward to future steps to determine if there is a better overall solution. The present minimum set-cover approximation uses a greedy solution to minimize the number of nodes, in this case channels, to select to reduce the total number of channels that need to be reviewed to "cover" a maximum number of channels in the graph. In particular, the algorithm iterates through selecting a candidate channel that is related to the most channels that need at least one more related spam channel to be suspended.

Figure 1 depicts a flow diagram illustrating an example method 100 for applying a minimum set-cover approximation algorithm to a similarity graph. At step 102, the algorithm may begin by counting how many related spam channels each channel in the graph needs to be connected to in order to be suspended. As discussed above, a channel that hasn't been reviewed but is related to other spam channels may be identified as a spam channel if it is related to a specified number "N" of spam channels that have been identified as spam. The number of spam channels that a channel must be related to in order to suspend the channel may be chosen to provide a very high probability (e.g., 99% likelihood) that the channel is spam. Thus, the algorithm may determine how many spam channels each channel is connected to and then calculate how many more spam channels that the channel must be connected to in order to identify the channel as spam. In one example, the edges of the similarity graph are weighted based on the extent of similarity between channels. If the edges are weighted, then edge weight may be taken into account when identifying spam channels based on connections. In some instances, rather than a specified number of connections, a threshold total weight of connections may be used to identify a channel as spam.

At step 104, for each candidate channel for human review, determine how many related channels it has that need at least one more spam review. A candidate channel is a channel that is a present candidate for human review. Each candidate channel may be connected to some number of channels that would need to be connected to at least one more spam reviewed channel in order to be suspended. Thus, the candidate channels may be associated with one or more related channels that are not connected with enough other identified spam channels to be accurately identified as a spam channel. These channels that a not related to N number of spam channels (i.e., cannot accurately be identified as spam) are counted for each candidate channel.

At step 106, select the candidate channel with the largest number of related possible spam channels, then add this channel to the set of channels to send for review. The channel that has the largest number of related possible spam channels may have the largest potential to identify a larger number of spam channels. The selected candidate channel may be added to a list of selected channels to be human reviewed.

At step 108, the algorithm decrements the number of related spam channels needed for all related channels. In other words, all channels connected to the selected channel require one less related spam channel because it is assumed that the selected channel will be identified as spam. At step 110, the previous three steps are then repeated until a threshold number of channels have been selected. At step 112, once the threshold number of channels are selected then the selected channels may be provided to human reviewers for review.

Figure 2 depicts a similarity graph of channels that represents the similarity between channels through edges connecting similar channels. Some channels may already be identified as spam channels (e.g., S1 and S2). Some set of other channels may be potential spam channels that may or may not be connected to other spam channels (e.g., C1, C2, C3, and C4). Other channels may be candidates for human review (e.g., H1, H2, and H3).

Under the first step of the algorithm described in Fig. 1, the number of spam channels connected to each channel (C1, C2, C3, C4) is calculated. For example, the necessary number of related channels in the present example may be N=2. As depicted in Fig. 2, C1 is connected to S1 and S2 and thus does not need to be connected to any more spam channels to be suspended. C2 is connected to S1 and thus needs only be connected to one more spam channel to be suspended. C3 is connected to S2 and thus needs only be connected to one more spam channel to be suspended. C4 is not connected to any spam channels so it needs to be connected to two more spam channels to be suspended.

Under the second step, the number of related channels that need at least one more review for each candidate channel for human review is calculated. In this example, H1 is connected to C2, C3, and C4. H2 is connected to C4 and H3 is connected to C2 and C3. Under the third step, H1 is selected because it has the most related channels that need at least one more spam reviews. C2, C3, and C4 each need at least one more connected spam channel, and therefore H1 has three related channels that need at least one more spam review. H2 is connected to one channel, C2, which needs one more connected spam channel. H3 is connected to two channels that need more connected spam channels, C2 and C3. Thus, H1 has the most related channels that need more connections to identified spam channels. H1 may be added to a list of channels to be sent to human reviewer for spam review.

Once H1 is selected, each of the channels related to H1 need one less spam channel for review. Therefore, C2 and C3 now need no more related spam channels, and C4 needs only one more related spam channels. Now the algorithm may be repeated with the remaining channels and the updated connections.

In the next iteration of the algorithm, in step one the channels C1, C2, and C3 do not need any more related spam channels. C4 still needs one more related spam channel. Then, in step two H2 is connected to one channel that still needs at least one more related spam channel (i.e., C4). H3 is connected to C2 and C3. C2 and C3 are now connected to N=2 spam channels. Accordingly, H3 is not connected to any channels that need more related spam channels to be suspended. Therefore, H2 has the most connected channels that need more related spam channels and is selected to be added to the list of channels to be reviewed. As a result, the channels H1 and H2 are selected as the two channels that will cover the entire set of channels if reviewed. Channels H1 and H2 are then sent to reviewers so that they be reviewed by human reviewers.

The use of the minimum set-cover approximation algorithm to find channels to review for spam provides the advantage of reducing repetitive and wasteful human reviews. The application of the algorithm may provide for fewer human reviews resulting in a larger number of identified spam channels. In addition, the use of the set-cover approximation algorithm may be easy to implement on a similarity graph already generated for a set of channels.

Further to the description above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein may enable collection of user information (e.g., information about a user's activities, information about content of documents, a user's preferences, or a user's current location), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

ABSTRACT

A method for selecting potential spam channels for human review by applying a minimum set-cover approximation algorithm to a similarity graph of channels. The nodes of the similarity graph represent channels and the edges between nodes represent similarity between channels. Channels may be reviewed by human reviewers to determine if they are spam. If a channel is connected to a certain number of identified spam channels then they can also be identified as spam channels and suspended. The algorithm may be an iterative process of selecting a channel to review that is connected to the most number of other channels that require another connected spam channel to be suspended.

Keywords: media content platform; minimum set-cover algorithm; spam; spam removal; spam channel identification; greedy algorithm; spam review

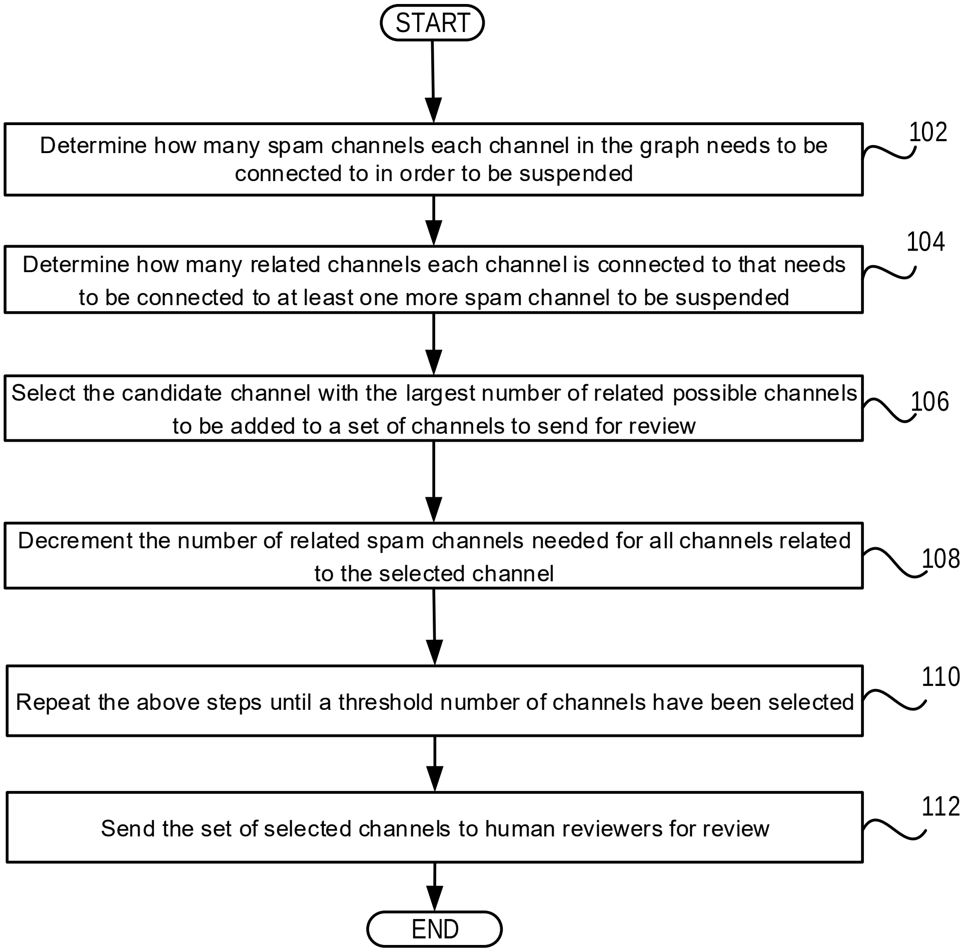
START
Determine how many spam channels each channel in the graph needs to be connected to in order to be suspended
102
Determine how many related channels each channel is connected to that needs to be connected to at least one more spam channel to be suspended
104
Select the candidate channel with the largest number of related possible channels to be added to a set of channels to send for review
106
Decrement the number of related spam channels needed for all channels related to the selected channel
108
Repeat the above steps until a threshold number of channels have been selected
110
Send the set of selected channels to human reviewers for review
112
END

FIG. 1

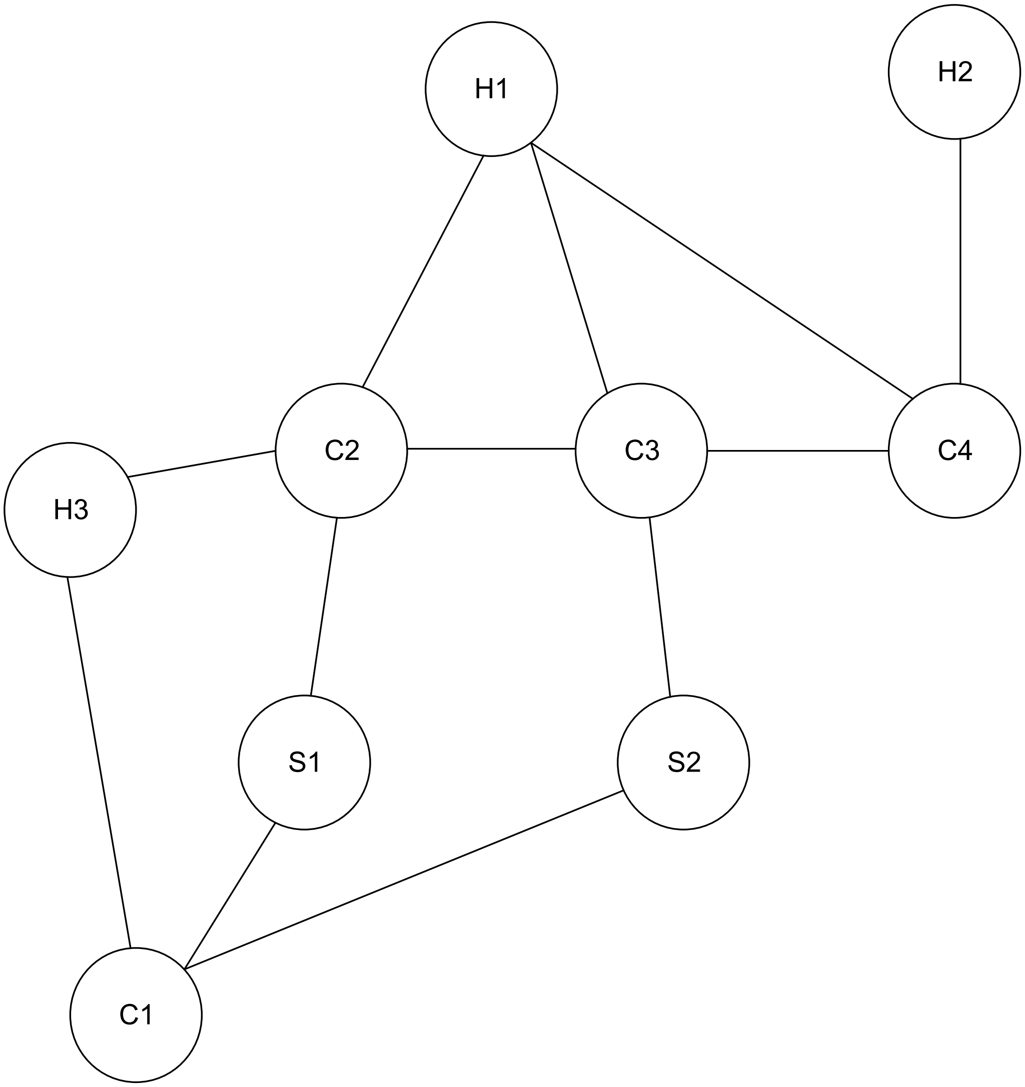

FIG. 2