Dynamic Surveillance: A Case Study with Enron Email Data Set

Heesung Do, Peter Choi, and Heejo Lee

Wantreez Music Inc. Seoul, Korea
Emulex Corporation, Costa Mesa, CA, USA
Div. of Computer and Communication Engineering
heesung@wantreez.com
peter.bkchoi@emulex.com
heejo@korea.ac.kr

Abstract. Surveillance is a critical measure to break anonymity. While surveillance with unlimited resources is often assumed as a means, against which, to design stronger anonymity algorithms, this paper addresses the general impact of limited resource on surveillance efficiency. The general impact of limited resource on identifying a hidden group is experimentally studied; the task of identification is only done by following communications between suspects, i.e., the information of whos talking to whom. The surveillance uses simple but intuitive algorithms to return more intelligence with limited resource. The surveillance subject used in this work is the publicly available Enron email data set, an actual trace of human interaction. The initial expectation was that, even with limited resource, intuitive surveillance algorithms would return the higher intelligence than a random approach by exploiting the general properties of power law-style communication map. To the contrary, the impact of limited resource was found large to the extent that intuitive algorithms do not return significantly higher intelligence than a random approach.

Keywords: Surveillance, Budget, Anonymity, Email Data Set

1 Introduction

One of the popular models of observer in the anonymity research is the one with unlimited resource and computing power such that the observer can monitor every single communication occurrence between any entities and exploit any possible derived information from the observation. Anonymity algorithms that can confuse such a powerful observer are regarded highly effective.

To understand the mighty power of the observer from a different perspective, one can ask this simple question, "what happens with limited resource?" This is the motivation of this paper. However, a glimpse of the anonymity research reveals the vast space of exploration to answer the question in a comprehensive manner. Different anonymity systems will cause different impact on resource-limited surveillance. This paper takes one small step to obtain insights into the impact of limited resource on surveillance.

The model for this work involves a simple anonymity group and simple surveillance algorithms; the anonymity group is the target for the surveillance to find. The target group does not employ sophisticated anonymity algorithms but encryption. The surveillance uses only the information of communication relationship (whos talking to whom) to find the entire target group.

The limited resource can be implemented in many different ways. In this paper, it is represented as the "budget", which is loosely defined as the unit of resource to monitor one subject (potential or identified hidden group member). So the number of subjects under surveillance is linearly proportional to the budget.

One consequence with the budget is that the surveillance has to make a decision at some points whether to continue to monitor the subjects currently under surveillance or replace the subject with another potentially more promising one. By "promising" it is meant that the new subject would likely be to reveal more members of the hidden group. Note that surveillance with unlimited resource would not need to change the monitoring subject. That kind of surveillance would just keep adding more subjects. This is the point where the attribute "dynamic" is introduced to better characterize the nature of surveillance with limited budget; the critical decision is made dynamically at points in time.

This dynamism creates the two parameters; period and selection algorithms. The period is some time amount, at the end of which, the surveillance makes a strategic decision to select more promising subjects for next surveillance period. The selection algorithms assign a priority to each candidate subject. Top priority subjects, as many as the budget allows, will be selected for next surveillance period.

The selection algorithms in this experiment are high-degree-first (HDF), high-traffic-first (HTF), and random (RAND). The "degree" means the number of edges from the node in the communication map. There is one-to-one relationship between the node in the communication map and one subject in the real world. The HDF assigns priority based on the degree; higher degree receives higher priority. Likewise, in HTF, higher traffic (higher communication occurrences) nodes receive higher priorities. Lastly, the RAND assigns priority in a random fashion. It is chosen as the baseline against which the performances of HDF and HTF are compared.

This paper uses the publicly available email data set of once American energy company Enron, as a trace of actual human communication. The process of identifying the target group is performed by following the communication of a selected target. The experiments show the general impact of limited resource on the intelligence obtained by the surveillance. The intelligence is measured by the number of Enron employees as the hidden group and the number of third parties who have communicated with any employee of Enron.

With the well known power-law phenomenon in social graph, where a few nodes are connected to a large portion of the entire nodes while a large portion of the nodes is connected only to a few other nodes, it may be natural to expect a maximum intelligence return from the surveillance by following the largest degree or largest traffic volume subjects.

Surprisingly, the simulated surveillance shows the opposite. Both HDF and HTF do not return noticeably higher intelligence than RAND. In other words, the impact of limited resource can be larger than expected.

The paper is organized as follows. A brief survey on related work and background information are given in Section II. The surveillance model, simulation overview and simulation data are described in Section III. Section IV details the impact of limited resource by showing the returning intelligence from HDF, HTF, and RAND with various periods and budgets. The concluding remarks and future work are provided in Section V.

2 Related Work

In a broad sense, this paper belongs to the other side of the general idea of anonymity research (for example, [2] [3] [9] [13] [16]). While the general goal of anonymity research is to hide the communication relationships and the participants identities, the goal of surveillance is to reveal such information.

There is one research work addressing the efficiency of surveillance at an abstract level [7]. The focus of the work of [7] is different from that of this paper, however. The former investigates the impact that the revelation of one single member of the target group brings to the discovery of the entire target group. Surprisingly, one single member revelation is found to divulge about 50 other members of the same target group. So, carefully planned surveillance would need to monitor only one fiftieth of the estimated target group population.

This paper, in comparison, treats each target subject individually. It does not take the clustering coefficient (relationships existing among members of the same group) into account. From the perspective of [7], this work can be said to investigate an extreme case, where each and every group has only one member. From some distance, this work seems to be related to the existing surveillance systems such as Carnivore [17] and NarusInsight [12]. The details of the systems are not known to the authors, however.

A number of research works have utilized the publicly available Enron email data set. Shetty et al. [14] created a MySql database from raw Enron email corpus, analyzed the statistics of the data set, and derived a social network graph. Keila et al. [11] explored the structure of the data set and analyzed the relationships among individuals by using the word use frequency. In addition to the study of analyzing the Enron email data set itself, some work [1][4][15] use the data set as a testbed for the applied research. In relation to this paper, one of them investigates the communication map of the email data set in great detail [8]. To the best of the authors knowledge, this paper is the first attempt to utilize the data set to investigate surveillance efficiency issues with limited resource.

3 Surveillance Model

3.1 Simulation Overview

The goal of this paper is to obtain insights into the impact of limited resource on the intelligence returned by surveillance. The intelligence in this experiment is to identify firstly the target (hidden) group and secondly the group of third parties who have communicated with any surveillance subject.

The target group is assumed unaware of the surveillance. It does not take any measure against the surveillance. So, whatever seen by the surveillance is the actual communication in this model.

The process of identifying the target group is performed by following the communication map drawn from observation of communication between one known subject and another subject. The content of the communication is assumed properly encrypted so that decipherment of the message is not practical. However, the identity of communicating subject is assumed to be decode-able by some means.

Since the surveillance finds more unidentified subjects anyway as time progresses, the communication map grows accordingly. However, the communication map adds only newly identified subjects. Otherwise, it adds more edges or increase traffic volumes. As the resource is limited, the communication map is always a subset of what has happened in the real world.

At the end of each monitoring time window, within the limited budget, the surveillance has to make a decision about which discovered subjects will be under next round surveillance. The selected subjects will determine the quality of next round surveillance because any new discovery will be done by identified communications with any of those subjects. The three algorithms for the target selection in this work are HDF, HTF, and RAND.

By identifying each subject this way the surveillance will eventually identify and establish the entire target group if time and budget allow. The simulated surveillance is done when the communication is exhausted, i.e., all the input data is exhausted. Different intelligence will be returned at the end of one simulation run with a different set of period, selection algorithms, and budget. This surveillance process is simulated by the software designed for the purposes.

To obtain one point in the figures in what follows the simulation is performed as follows.

1. A simulation data set is given, which is a trace of actual human interactions.
   (a) Each communication occurrence of the data set is associated with the time of occurrence and the sender and receiver.
   (b) so, the entire data set is a collection of communications on the time line from the beginning to the ending time points.
2. Set the surveillance period, subject selection algorithm, budget.
3. Read the first time slice of the simulation data set based on the period.
4. At the beginning of the first period,
   (a) Select some subjects from the slice randomly as many as the budget.

   (b) Put those under surveillance.
   (c) Run surveillance.
      i. Observe the communication chronologically.
      ii. Create the communication map accordingly.
5. At the end of the first period, run the subject selection algorithm.
   (a) Select the top priority subjects as many as the budget.
   (b) Put those under surveillance.
6. Read the next time slice of the simulation data set based on the period.
   (a) Run the surveillance with the subjects selected at the end of the previous round.
      i. Observe the communications with the selected subjects chronologically.
      ii. Update the communication map accordingly.
7. At the end of the current period, run the subject selection algorithm.
   (a) Select the top priority subjects as many as the budget.
   (b) Put those under surveillance.
8. Repeat the above two steps (6, 7) until the input data set is exhausted.
9. At the end of the run report the intelligence.
   (a) The identified subjects of the hidden group.
   (b) The identified third party subjects, who have communicated with one of the identified subject of the hidden group.
   (c) Other information as desired.
10. Repeat the entire procedure above 30 times with the same set of period, selection algorithm, and budget.
11. Obtain the averaged intelligence of the 30 runs.

The averaged intelligence should not be affected by the seed subjects, which are randomly selected from the first time slice of the simulation data set.

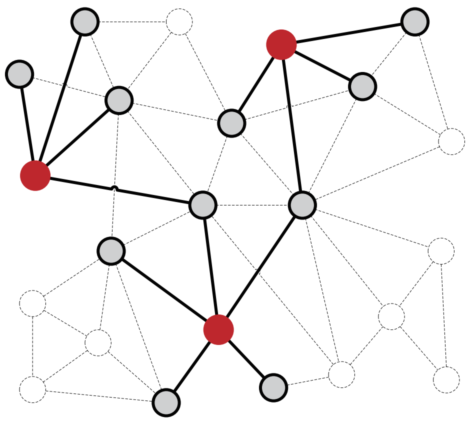


Fig. 1. Illustration of the surveillance model under limited budget. This example shows budget 3, so that only three nodes (3 red nodes) can be put under surveillance.

Illustration of Limited Resource Figure 1 shows an example of surveillance under a limited budget. The graph represents a communication map among subjects from a period. The budget is set to 3. There are 23 nodes in the communication map. However, the surveillance can only identify 14 nodes; 3 red (or dark in black-and-white) nodes and 11 grey nodes. The rest 9 nodes cannot be observed by the surveillance due to the limited budget. In other words, the surveillance returns the intelligence of the 11 discovered nodes.

3.2 Simulation Data

The input data to the simulation is the Enron email data set. So, in this work, each unique email address is treated as an unique individual or a possible surveillance subject. The target group is the set of unique email addresses which are in the form of "somename@enron.com". Identifying the target group then becomes identifying all unique email addresses which end with "@enron.com". The third parties are identified when their communication with any known subject is identified by the surveillance

The first public release of the Enron email data set was done in May 2002 by the Federal Energy Regulatory Commission [6]. Since the public release, several groups have subsequently processed and used the data set for a range of different research purposes. As a result, there are a few different versions available now. In this paper, the ISI (Information Sciences Institute) MySql version [10] of the data set is used. The ISI version was originally based on the CMU (Carnegie Mellon University) version [5].

The CMU version contains 517,431 messages from 151 employees. By removing meaningless messages from the CMU version, the ISI version now holds 252,759 messages from 151 employees, about half of the CMU version. This work slightly improves the ISI version in terms of message validity for surveil- lance purposes. As a result, the MySql file size changes from 740 Mbytes (ISI version) to 667 Mbytes in this work. The data set used in this work has 252,692 email messages, 75,529 unique email addresses from Jan. 4, 1998 through Dec. 21, 2002.

So, the simulated surveillance is to identify all the 151 employees (someone@enron.com) and other third parties who communicated with one of the employees. Figure 2 shows the message distribution for the 5-year time period. The message volume peaks around Oct. 2001.

4 Experimental Results

4.1 Simulation Specifics

Surveillance time window The two types of window are used in this work; time based and message based. In the time based, the entire data set is divided by a time period. Each window has the same time span. Some windows see a large number of email messages while some others do not. In the message based,

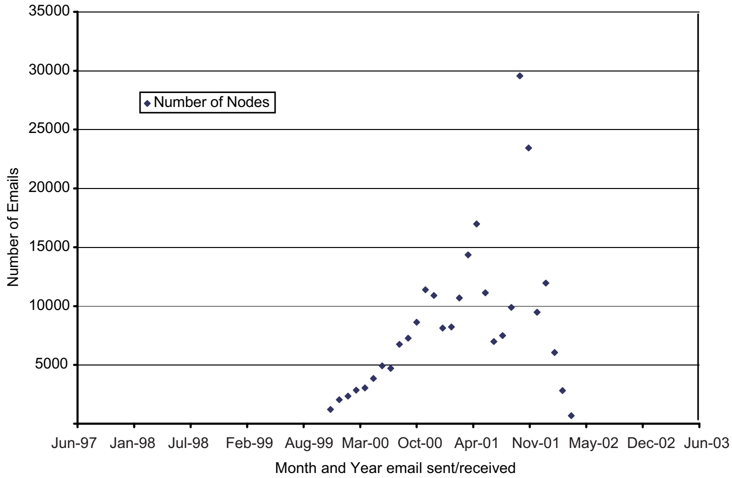


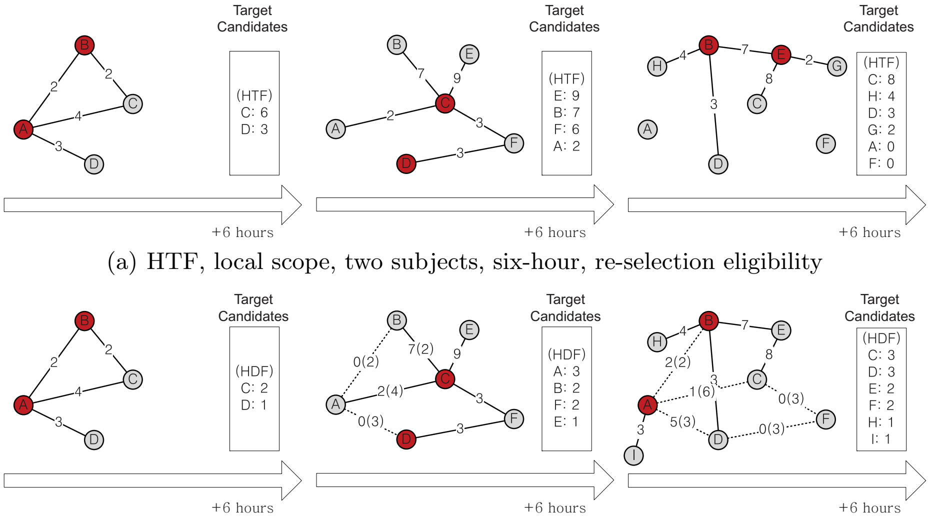
Fig. 2. Distribution of email messages in time [14]

the entire data set is divided by a number of messages. Each window has the same number of messages. Some windows take a large time span while some others take a short time span. Once the simulation started, each window is fetched from the MySql database in sequence, and, is given to the simulation software for surveillance processing.

Target selection scope The target nodes in the communication map are the subjects, with which the simulated surveillance runs for the next time window. This work uses three simple strategies for target selection; HDF, HTF, and RAND at the end of each surveillance period. In the process of target selection, the surveillance needs to see the pool of candidates. The pool can be formed in two ways; local and global. The local pool is formed by the nodes observed in the current window of communication map. The global pool is formed by the entire nodes observed from the beginning up to the current window included. The local pool has fewer candidates while the global pool has an increasingly large number of candidates. Depending on the setting of the target selection, the three strategies (HDF, HTF, and RAND) select the target nodes either from the local or from the global scope.

Eligibility of Re-Selection

1. *Rule 1:*
   A target node in this work is not allowed to be selected again to be one of the target nodes for the immediately following window. A target node however can be selected again as one of the target nodes for the window, which is at least two window-hops away. Two neighboring windows are one window-hop away from each other. This is different from the target selection scope. The



(a) HTF, local scope, two subjects, six-hour, re-selection eligibility

(b) HDF, global scope, two subjects, six-hour, eligibility

Fig. 3. Illustration of surveillance example (target selection strategy, target selection scope, budget, window, re-selection eligibility)

target selection scope defines the pool of candidates. This rule defines the eligibility of re-selection.

2. *Rule 2:*
One assumption of the simulation is that the target nodes will communicate with unknown parties during the next surveillance window so that the surveillance will identify more suspect nodes. A question arises when the target nodes do not exhibit communication with any unknown parties. The two choices are available in this case for selecting target nodes;
   (a) Select one of the target nodes from the past for the next window, or,
   (b) Use the target nodes of the current window for the next window because no new candidates are available from the current window.

This work uses the second choice. This "continued status of the target node over two neighboring windows" is one exceptional case to the "Rule 1" of eligibility of re-selection.

Illustration of Simulation Specifics Figure 3 shows an illustration how the surveillance works with different target selection strategies and scopes with the same simulation data set for the same time windows and the same re-selection eligibility rules; (a) shows HTF in the local scope in the three consecutive windows, (b) shows HDF in the global scope in the same three consecutive windows. The setting includes a time window of 6 hours, budget 2. The red (or dark in black-and-white) nodes are the target nodes. The edge represents the identified communication between nodes. The weight of the edge is the communication volume; the number of email messages exchanged by the pair of nodes.

In the first window of the Figure 3 (a), there are two target nodes (A, B). Through the target nodes, the surveillance observes the communication between A and B, A and C, A and D, and, B and C. In terms of HTF, A is still the highest traffic node with 9 communications. However, since there are other unknown nodes, C and D, A is not allowed to be selected to be a target node for the next window. Accordingly, C and D are selected as the target nodes for the next window of 6 hours. In the second window of (a), C is the highest traffic node. Again, however, other new so-far unknown nodes are selected as the target nodes for the third window; E and B.

The same rules are applied in Figure 3 (b). The differences are the target selection strategy and scope; HDF in the global scope. The target nodes of the first window are A and B. At the end of the first window, A and B are still the highest degree nodes. Due to the "eligibility of re-selection", however, C and D are selected as the target nodes for the second window. The global scope of the second window is represented by the solid and dotted lines between nodes. The solid line is the communication occurred in the current window. The dotted line is the communication observed in previous windows. Likewise, the number in the parenthesis on the edge is the cumulative communications between the pair of nodes up to the immediately preceding window, while the number out of the parenthesis represents the communications observed in the current window.

In Figure 3 (b), both B and F have the same degree, 2, at the end of the second window. The tie is broken in this work in favor of higher traffic; B has 7(2) + 0(2), while F has 3 + 3. Eventually A and B are selected as the target nodes for the third window. Note that A and B were the target nodes for the first window. Both A and B are eligible to be a target node for third window because the first and third windows are two window-hops apart. Note that the two communication maps made by HDF and HTF grow differently with the same simulation data set. Both communication maps are incomplete anyway due to the limited resource.

4.2 Dynamic Surveillance with Limited Budget

In the figures below, the "suspects"are the unique email addresses of 151 Enron employees. The "nodes" are the unique addresses, which are either suspects or any other addresses, which have communicated with the employees at least once during the surveillance.

Figure 4 shows six graphs which differ from each other in the window type and target selection scope. The first column (a, c, e) shows the node discovery, and, the second column (b, d, f) shows the suspect discovery. The first 4 graphs (a, b, c, d) are obtained using the target selection from the local scope, while the last two (e, f) are obtained by the global scope target selection. The X-axis shows the surveillance window size in either message or time and its corresponding percentage of the entire surveillance period (5 years). Note that the X-axis is not a time line. One simulation run produces a value at one point of the curve. The

Y-axis show the averaged intelligence either node or suspect discovery percentage against the entire data set size with the given budget, window, target selection strategy, and target selection scope.

The first two (a, b) use time windows while the last four (c, d, e, f) use message windows. Each graph has five sets of curves; each set represents the budget 4, 16, 64, 256, and 1024. Each set of the graph in turn shows the performance of the three target selection strategies; HDF, HTF and RAND under the same conditions of budget and window. Each graph has fifteen curves in total, therefore.

Each point is obtained by the average value of 30 simulation runs with the same simulation setting but different seed nodes. For example, in (a), both HDF and HTF with the window of 96 hours and budget 256 produce the node discovery ratio of about 25%. This is the averaged value of 30 simulation runs. So each graph is a collection of averaged values from a set of independent simulation runs. Simulation runs higher than 30 do not produce noticeable difference. The best possible node discovery in this experiment as seen in the figure is about 35% or 36% of the entire nodes when the email data set is exhausted.

Global vs. Local Scopes The last two graphs (e, f) use the global target selection scope while the first four (a, b, c, d) graphs use the local scope. One can expect that the global scope would return higher intelligence because the larger pool of candidates. To the contrary, the results are the opposite. The node discovery rates of (e) are lower than those of (a) and (c). Similarly, the performance of (f) is lower than (b) and (d). The reason is in the limited budget. The global scope tend to select the same target nodes again in later windows due to their accumulated higher degrees and traffic volumes. This trend prevents other new more promising nodes from being selected. The local scope, however, has to select the target nodes from the new local pool at each window.

Budget vs. Discovery Rates With the increasing budgets, the 151 suspect nodes (employee addresses) are 100% discovered. As can be seen from (b), (d) and (f), the complete suspect (employee) discovery is achieved with the budgets 256 and 1,024. So, budgets higher than 1,024 are not experimented. The graphs (a), (c) and (e) show that higher budgets yield higher discovery of nodes. However, while the budget is increased by 4 times at each step, the discovery ratio increases only sub-linearly.

The ratio of discovery to budget is found only to decrease. With this kind of sub-linearity, an absurdly large budget would be required to discover higher nodes than shown in (a) and (c). Further, the return intelligence is found increasingly marginal from each multiplicatively higher budget investment.

Time vs. Message-BasedWindow In this experiment, as can be seen in Figure 4 (a) and (c) or (b) and (d), no big performance difference is found between the two different kinds of surveillance period; time and message windows. This is somewhat counter intuitive because the number of communication occurrences in the time window is likely to be different for each period. The logical explanation to this is that the variation of the message volume in the time window was not to the extent, where performance degradation would be seen. As seen later, both windows find new nodes at a rather constant rate.

HDF, HTF vs. RAND In (a) and (c), the set of curves seems to have a mild peak. Interestingly, the three selection strategies do not show much performance difference until that point. After the peak, RAND shows the lowest performance while HTF is only slightly lower than HDF. Throughout the range of budgets, HDF and HTF do not show noticeable difference. One possible logical explanation to these results is that, up to some window sizes and budgets, for example, 512 messages or 48 to 96 hours and 64 or higher budgets, intuitive algorithms do not necessarily perform better than a random approach. In other words, the windows and budgets up to the peak point may not be large enough for the intuitive algorithms to exploit some patterns in the communication maps.

Peak Interestingly, in (a) and (c), there tends to be a peak in the node discovery ratio. For example, in (a), the node discovery reaches about a little more than 35% with the budget 1024, the window of 48 hours regardless of the strategy. Similarly, in (c), the ratio reaches about 36% with the budget 1024, the window of 512 messages, again, regardless of the strategies. The peak becomes more recognizable with higher budgets. In this work, the peak is interpreted that budgets larger than certain percentage of the entire nodes may have some optimal range of windows to maximize the return intelligence.

The peak is clearer in the message windows in (c) although the overall performances are not much different from those of time windows in (a). This is because the number of message appearing in each window is constant in (c), while it is necessarily fluctuating in the time windows in (a). The even distribution of messages in (b) must have helped manifest the optimal range of windows.

The performance degradation of (a) and (c) after the peak point is also interpreted due to the larger window. The peak point is effectively the turning point where the window size becomes sufficiently large to create the global scope effect for target selection. By the same argument, the global scope also produces flat curves in (e),

Another side effect of the global scope is the larger gap between RAND and the other two (HDF, HTF) with large budgets (256, 1,024). In (a) and (c), the gap between RAND and the other two becomes visible only with large budgets and large windows. Statistically RAND has higher probability to choose worse nodes in the global scope than in the local scope. The wider variety of the global scope contributes to the poor target node selection of RAND. In the local scope, since it is always created by the most promising nodes from the previous window, RAND has lower probability to choose low performing nodes.

4.3 Variations of Dynamic Surveillance

Strategically Uneven Budget Allocation So far, the budget is evenly allocated to each window. This is to reflect the general situation that the dynamic surveillance would not know when more new nodes would appear in the surveillance. Without knowing the future information, the strategy of even budget allocation would be a reasonable choice.

The general question is whether there would be a better way of budget allocation in an effort to improve node discovery. To be fair, the total amount of budget needs to be assumed fixed. The total amount of budget is defined as the average budget per window multiplied by the number of windows of the entire surveillance period, 5 years.

One immediate way is to allocate a relatively large portion of budget to the early stage of surveillance. The idea is to exploit the general pattern of communication map that a small percentage of nodes are connected to most of the nodes.

The hope is that if such small percentage of nodes would be discovered at an early stage, the node discovery would be more effective for the rest of the surveillance even with less amount of budget to the following windows. Therefore, the two variations of budget allocation are experimented here: firstly 50% of the total budget to the early 10% of the surveillance period, secondly 90% of the total budget to the early 10% of the surveillance period. The rest of surveillance windows receive the even distribution of the remaining budget in both cases.

Figure 5 shows the results of the two cases; (a) and (c) show the node and suspect discovery rates for the first case (50% allocation first), and, (b) and (d) show the second case (90% allocation first). In comparison to Figure 4 (c) and (d) (message window, local scope), the node discovery rates of Figure 5 (a) and (c) are not higher, and those of Figure 5 (b) and (d) are lower. These results apparently do not support the hope of finding more node.

More interestingly, in Figure 5, (b) and (d) (90% budget to the first 10% of surveillance period) show even lower rates than in (a) and (c) (50% budget to the first 10% of surveillance period). This result means that higher budget allocation to the early stage results in even lower node discovery. In an effort to understand this interesting result, the micro behavior of node discovery is further analyzed next.

Micro Observation of Node Discovery Figure 6 shows the ”progress” of node discovery of three budget allocation cases; even, 50% first, and 90% first allocations. The X-axis shows the time line in the number of surveillance windows. The Y-axis shows the return intelligence either the number of nodes identified (a, c, e) or the number of suspects (employees) (b, d, f) as the one time simulation progresses on the time line. As such, the returning intelligence (Y-axis) only grows on the time line (X-axis).

Note that these figures are different from the previous ones (Figure 4), where the curves show the averaged return intelligence of multiple independent simulation runs. Different points of the curve are from different simulation runs. In comparison, different curves of Figure 6 are from different simulation runs. The points of one curve are all from the same simulation runs.

The left column of three graphs, (a), (c) and (e), show the node discovery and the right three (b, d, f) show the suspect discovery. The first row, (a) and (b), are for the even distribution, the middle two (c) and (d) for the 50% first, and, the bottom two (e) and (f) are for the 90% first.

The highlight of this figure is the growing rate of the returning intelligence. In (a), the even distribution of budget, the node discovery grows almost linearly and eventually tops around 27,000 nodes, which is about 35% of the entire nodes.

In (c) the discovery grow rapidly for the first 10% surveillance period and the growth rate goes down immediately after the first 10% surveillance period. This phenomenon stands out more distinctively in (e). This trend remains the same even in the suspect discovery rates in (d) and (f).

Interestingly, in (e), the 90% first does not boost the node discovery rate even for the early 10% of surveillance period in comparison to (c). Evidently, this tells that more than 50% budget allocation to the early 10% of surveillance would not result any more intelligence return in this case study.

From a slightly different angle, this also suggests that the higher budget allocation to the early 10% of surveillance was not much effective because the possible pattern (power-law, for example) of communication map was not fully recognizable in the early stage even by the temporarily large budgets. So, in this case study, choosing the even budget distribution seems favorable for the two selection algorithms, HDF and HTF.

4.4 Surveillance with Unlimited Resource

Using the same simulation data set, this section runs the simulation with unlimited resource, i.e., the surveillance monitors every single communication occurrence between any two nodes. The communication map is complete at any given moment, therefore. The motivation is to see the difference between the intelligence returned by resource-limited and -unlimited surveillance.

Figure 7 shows four graphs on the X-Y plane with a logarithmic scale on the X axis. As before, the Y-value is the ratio of node (unique email address: both employee and third party combined) discovery. The X axis shows the top percentage of nodes with the priorities assigned by the target selection algorithm.

For example, in (b), the top 1% of nodes selected (on X-axis) either by HDF or HTF are connected with the other 70% or higher (Y-axis) nodes of the communication map. This means that the selection of top 1% nodes by the selection algorithms can identify more than 70% of the nodes at the given moment. Since the surveillance has unlimited computing power, each single node or communication addition causes a new complete computation of the entire communication map. This allows the algorithm to assign the priority based on the exact global and complete view at any moment.

The four graphs are obtained as follows. First, take the first 0.1%, 1%, 10%, and 100% portion of the simulation data set from the time line. (Remember that the simulation data set is a chronologically ordered communication occurrences among subjects.) Second, sort out the selected portion using the three algorithms; HDF, HTF, and RAND. Here, all the nodes, which ever communicated with any of the selected nodes are considered discovered. Third, create a curve for each selection algorithm for the four different sets.

Since the four first portions (0.1%, 1%, 10%, and 100%) are different in size from each other, the connectivity of the top percentage of the first portion to the rest of the first portion is different from each other, too. For example, the node discovery by top 1% is more than 80% in (a), more than 70% in (b), more than 50% in (c), and lastly more than 40% in (d). The larger the first portion, the smaller the top percentage nodes connectivity.

Note that Figure 7 cannot be directly compared to Figure 4, where the X-axis was a time line while it is the top percentage of priority by the chosen target selection algorithm.

One convenient way to interpret the four graphs is to regard each one (a, b, c, d) as the snapshot of the surveillance with unlimited resource at the moments where the communication map reaches the first 0.1%, 1%, 10%, and eventually 100% of the entire nodes. Because it is resource limitation-free, the surveillance knows exactly what has happened. The current communication map itself reveals 100% discovery at any time. This is the big difference between the resource-limited and -unlimited surveillance.

With the always complete Communication map a few interesting observations are readily available.

1. As the surveillance progresses, HDF returns higher intelligence than HTF,
2. RAND returns constantly poor intelligence.
3. The curve patterns do not seem to change regardless of the size of the early portion of data set.

Considering these observations, it can be said that there maybe some patterns in the complete communication map, and, the HDF seems to exploit the patterns most effectively. It indirectly shows that the pattern may be a power law-style. Since RAND does not utilize any pattern, it should return the worst intelligence.

There is an interesting observation with the sizes of window. Figure 4 uses a range of window sizes. For example, the largest window size in Figure 4 (c) is 16,384 messages, which corresponds to about 6.5% of the entire data set. This window size is actually larger than those of Figure 7 (a) and (b). The largest window of Figure 4 (a) is 3,072 hours, corresponding to about 8.5% of the entire surveillance period. Interestingly even these large window sizes do not make the node discovery higher than 40% in Figure 4 (a) and (c).

Again, the major contributor to this interesting result is the incompleteness of the communication map due to the limited budget. The incomplete map constantly leads a sub-optimal selection of target nodes for next surveillance round. This phenomenon continues even with considerably large window sizes.

Lastly, the lowest curve in Figure 7 (d), is a hypothetical case, in which the target group uses an anonymity system such that the node discovery is perfectly linearly proportional to the surveillance budget. So, in order to find out X number of subjects of the target group, the budget of X should be invested. Finding the existence of such an anonymity system is out of the scope. This case, however, gives the lower bound to the surveillance performance. Even RAND performs better than this imaginary case.

5 Conclusions

The motivation of this work is to obtain insights into the impact of limited resource on the intelligence returned by surveillance. This work takes an experimental method in an effort to approach the right answer. The experiment was done in a form of simulated surveillance using a publicly available Enron email data set. The data set does not contain a complicated anonymity algorithms except data encryption. So the target selection algorithms were simple for the surveillance. However, the nature of the data set, a reflection of human interactions as a real trace, gives some credit on the actuality of the data set.

The experiment was done firstly with limited resource and followed by another form of surveillance with unlimited resource for comparison. As seen in the two strikingly different graphs (Figure 4, Figure 7), the impact of limited resource can be larger than expected. As seen in Figure 4, the idea of exploiting some intuitive patterns (high degree or high traffic) on the communication map was not effective with limited budgets. After the peak points, larger budgets and larger window sizes produced worse intelligence. Although both HDF and HTF perform much better that RAND after the peak, the intelligence returned by both was monotonically decreasing with considerably larger budgets and window sizes.

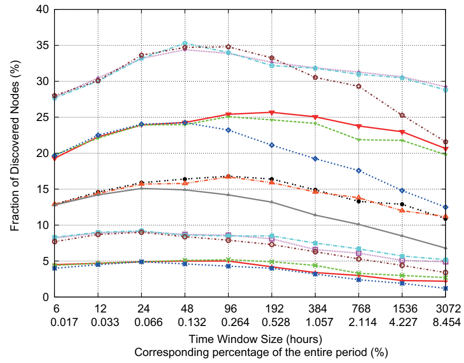
By comparing the two surveillance cases (resource limited vs. unlimited), even though this work is about only one single case study with Enron email data set, some conclusions can be drawn that:

- Surveillance with limited resource may have some optimal points in terms of the combination of budget and window size that can maximize the quality of intelligence returned by the surveillance.
- Even allocation of budgets throughout the surveillance may work better than strategically uneven allocations.
- The incompleteness of the communication map seems to be maintained throughout the surveillance. This may be the major contributor to the observation that both HDF and HTF do not return significantly higher intelligence than RAND.

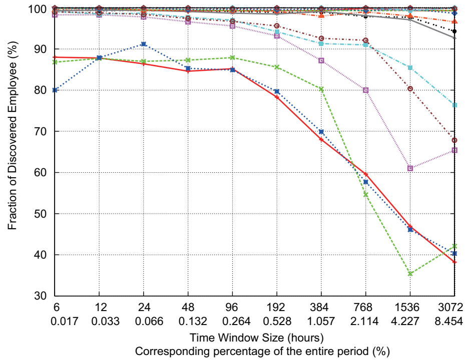
This work, although the generality is limited due to the scope of single case study, solicits further work, including but not limited to, on the optimal combination of budget and window size while the hidden group size is still unknown (with possible estimates of the group size), and, on the minimum size of communication map that is yet large enough to show some patterns to be utilized.

References

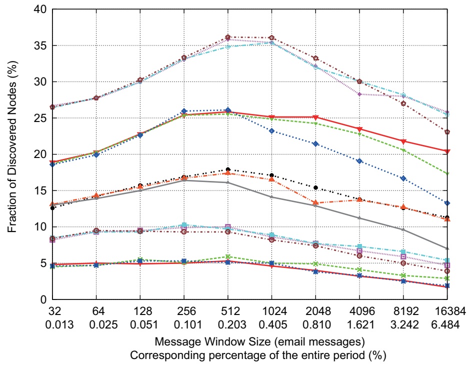
1. L. Akoglu, M. McGlohon, and C. Faloutsos. Anomaly detection in large graphs. In CMU-CS-09-173 Technical Report, 2009.
2. N. Bansod, A. Malgi, B. K. Choi, and J. Mayo. Muon: Epidemic based mutual anonymity in unstructured p2p networks. *Computer Networks*, 52(5):915–934, 2008.
3. O. Berthold, H. Federrath, and S. Köpsell. Web mixes: A system for anonymous and unobservable internet access. In *Workshop on Design Issues in Anonymity and Unobservability*, pages 115–129, 2000.
4. A. Chapanond, M. S. Krishnamoorthy, and B. Yener. Graph theoretic and spectral analysis of enron email data. *Computational & Mathematical Organization Theory*, 11(3):265–281, 2005.
5. CMU. Enron email dataset. http://www.cs.cmu.edu/ enron.
6. F. E. R. Commission. Enron investigation. http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp.
7. G. Danezis and B. Wittneben. The economics of mass surveillance and the questionable value of anonymous communications. In R. Anderson, editor, *Proceedings of the Fifth Workshop on the Economics of Information Security (WEIS 2006)*, Cambridge, UK, June 2006.
8. J. Diesner, T. L. Frantz, and K. M. Carley. Communication networks from the enron email corpus "it's always about the people. enron is no different". *Computational & Mathematical Organization Theory*, 11(3):201–228, 2005.
9. D. M. Goldschlag, M. G. Reed, and P. F. Syverson. Onion routing. *Commun. ACM*, 42(2):39–41, 1999.
10. ISI. Enron dataset. http://www.isi.edu/ adibi/Enron/Enron.htm.
11. P. S. Keila and D. B. Skillicorn. Structure in the enron email dataset. *Computational & Mathematical Organization Theory*, 11(3):183–199, 2005.
12. NarusInsight. Narusinsight solutions for traffic intelligence. http://www.narus.com/index.php/product.
13. R. Sherwood, B. Bhattacharjee, and A. Srinivasan. P^5: A protocol for scalable anonymous communication. *Journal of Computer Security*, 13(6):839–876, 2005.
14. J. Shetty and J. Adibi. The enron email dataset database schema and brief statistical report, 2004.
15. J. Shetty and J. Adibi. Discovering important nodes through graph entropy the case of enron email database. In *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05, pages 74–81, New York, NY, USA, 2005. ACM.
16. C. Shields and B. N. Levine. A protocol for anonymous communication over the internet. In *ACM Conference on Computer and Communications Security*, pages 33–42, 2000.
17. H. E. Ventura, J. M. Miller, and M. Deflem. Governmentality and the war on terror: Fbi project carnivore and the diffusion of disciplinary power. *Critical Criminology*, 13:55–70, 2005.



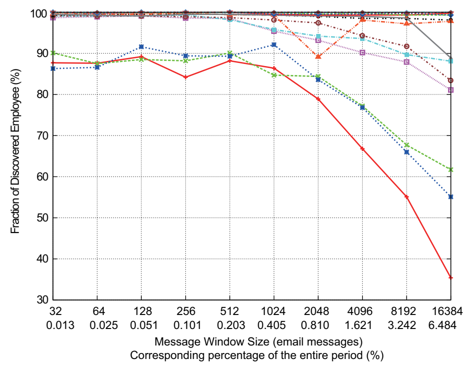
(a) Discovered nodes (TIME/LOCAL)



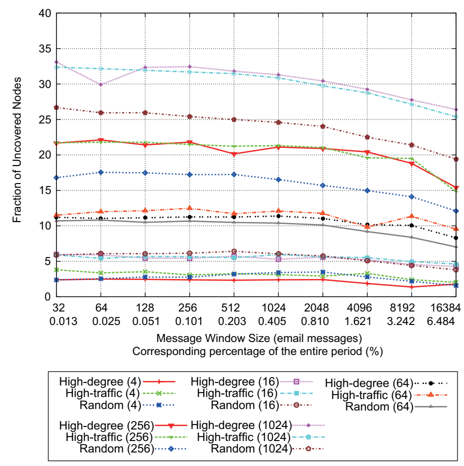
(b) Discovered employee (TIME/LOCAL)



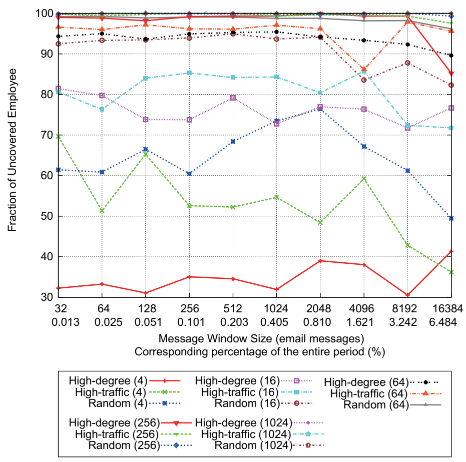
(c) Discovered nodes (MESSAGE/LOCAL)



(d) Discovered employee (MESSAGE/LOCAL)

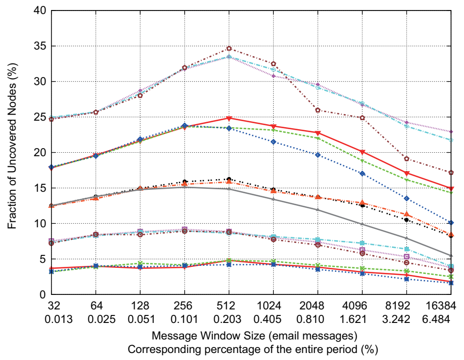


(e) Discovered nodes (MESSAGE/GLOBAL)

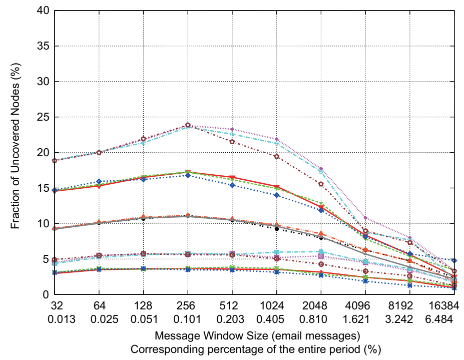


(f) Discovered employee (MESSAGE/GLOBAL)

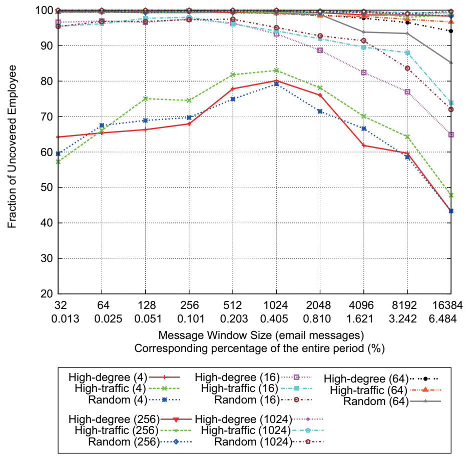
Fig. 4. Node discovery of dynamic surveillance



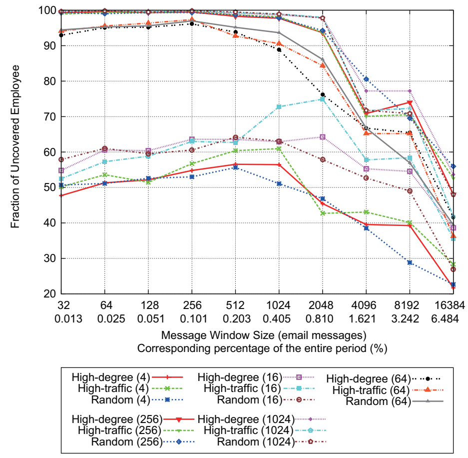
(a) Discovered nodes with 50% budget on early 10% windows



(b) Discovered nodes with 90% budget on early 10% windows

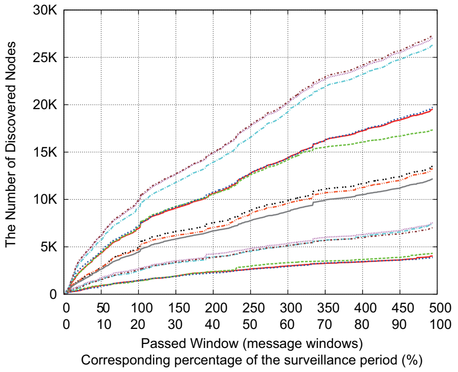


(c) Discovered employee with 50% budget on early 10% windows

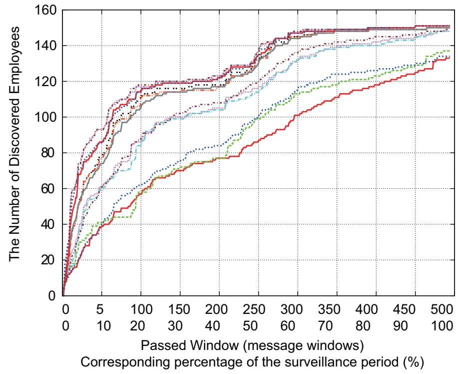


(d) Discovered employee with 90% budget on early 10% windows

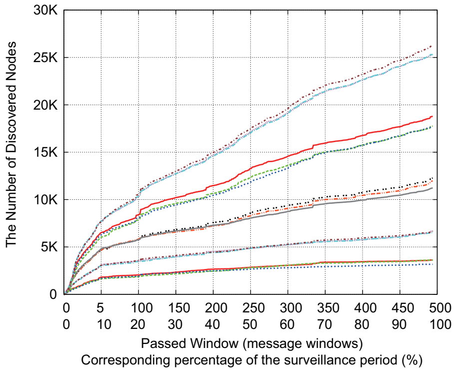
Fig. 5. Node discovery with variable budget distribution in before-event surveillance (message window based)



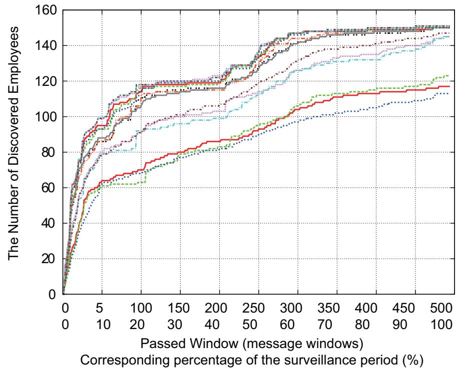
(a) Node discovery with evenly allocated budget



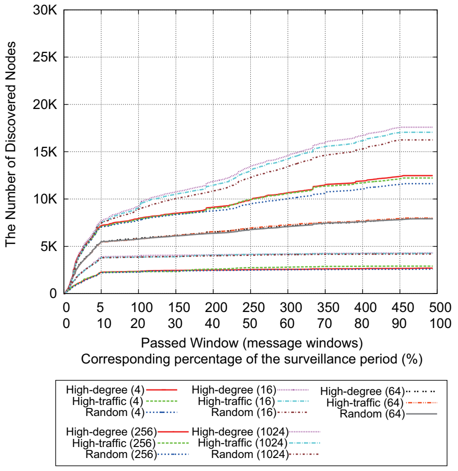
(b) Employee discovery with evenly allocated budget



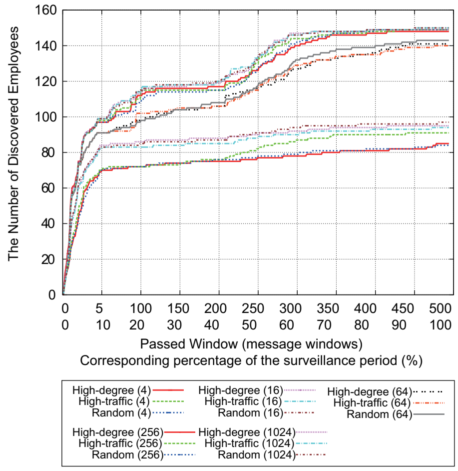
(c) Node discovery with 50%budget on early 10%windows



(d) Employee discovery with 50%budget on early 10%windows

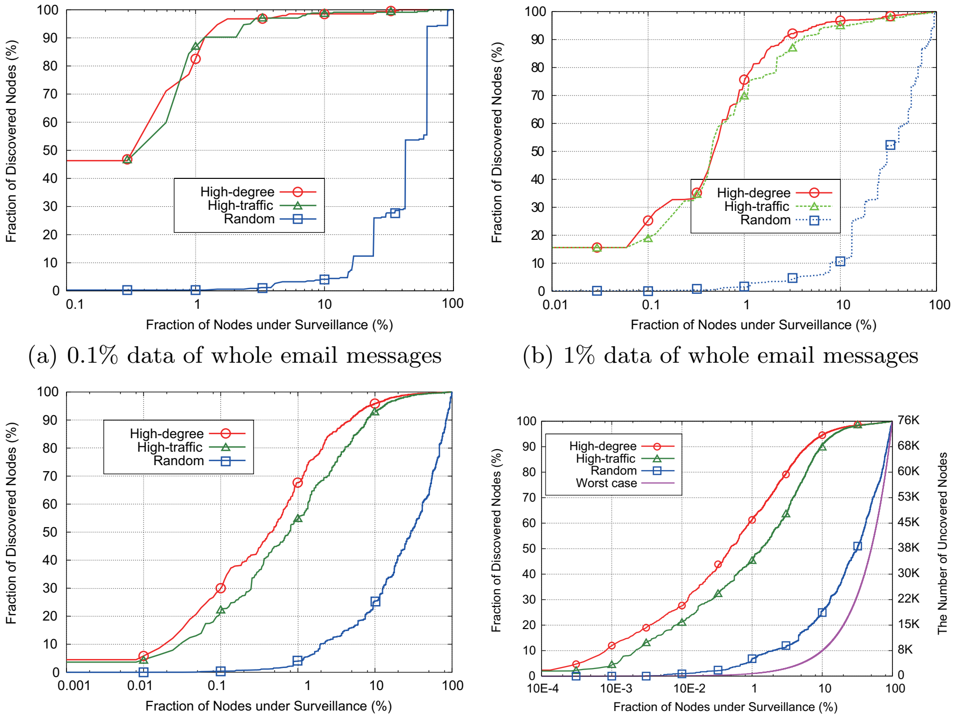


(e) Node discovery with 90%budget on early 10%windows



(f) Employee discovery with 90%budget on early 10%windows

Fig. 6. Progress of node discovery with various budget allocation cases



(a) 0.1% data of whole email messages

(b) 1% data of whole email messages

(c) 10% data of whole email messages

(d) 100% data of whole email messages

Fig. 7. Surveillance with unlimited resource