Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

# Node discovery in a networked organization

Yoshiharu Maeno
Social Design Group
Bunkyo-ku, Tokyo 112-0011, Japan.
maeno.yoshiharu@socialdesigngroup.com

***Abstract*—In this paper, I present a method to solve a node discovery problem in a networked organization. Covert nodes refer to the nodes which are not observable directly. They affect social interactions, but do not appear in the surveillance logs which record the participants of the social interactions. Discovering the covert nodes is defined as identifying the suspicious logs where the covert nodes would appear if the covert nodes became overt. A mathematical model is developed for the maximal likelihood estimation of the network behind the social interactions and for the identification of the suspicious logs. Precision, recall, and F measure characteristics are demonstrated with the dataset generated from a real organization and the computationally synthesized datasets. The performance is close to the theoretical limit for any covert nodes in the networks of any topologies and sizes if the ratio of the number of observation to the number of possible communication patterns is large.**

***Index Terms*—Anomaly detection, Covert node, Maximal likelihood estimation, Node discovery, Social network.**

## I. Introduction

Covert nodes in a networked organization refer to persons who affect social interactions (communications among the nodes and resulting collaborative activities), but do not appear in the surveillance logs which record the participants of the social interactions. They are not observable directly. Discovering the covert nodes is defined as identifying the suspicious surveillance logs where the covert nodes would appear if the covert nodes became overt. This problem is called a node discovery problem.

Where do we encounter such a problem? Globally networked clandestine organizations such as terrorists, criminals, or drug smugglers are great threat to the civilized societies. Terrorism attacks cause great economic, social and environmental damage. Active non-routine responses to the attacks are necessary as well as the damage recovery management. The short-term target of the responses is the arrest of the perpetrators. The long-term target of the responses is identifying and dismantling the covert organizational foundation which raises, encourages, and helps the perpetrators. The threat will be mitigated and eliminated by discovering covert leaders and critical conspirators of the clandestine organizations. The difficulty of such discovery lies in the limited capability of surveillance. Information on the leaders and critical conspirators are missing because it is usually hidden by the organization intentionally.

In this paper, I present a method to solve the node discovery problem. The method infers the network topology and probability parameters behind the social interactions (by use of the maximal likelihood estimation), applies an anomaly detection technique to the surveillance logs, and identifies the suspicious surveillance logs. III presents the method. IV introduces the dataset generated from a real organization and the computationally synthesized datasets for performance tests. V demonstrates the precision, recall, and F measure characteristics with the datasets.

## II. Related Work

The social network analysis is a study of social structures made of nodes which are linked by one or more specific types of relationship. Examples of the relationship are influence transmission in communication, or presence of trust in collaboration. Studies in complex networks [3], [24], [6], WWW search and analysis [4], [1], and machine learning of latent variables [21], [22] are major related research topics.

Research interests have been moving from describing organizational nature to discovering unknown phenomena. A link discovery predicts the existence of an unknown link between two nodes from the information on the known attributes of the nodes and the known links [5], [7], [23]. The link discovery techniques are combined with domain-specific heuristics. The collaboration between scientists can be predicted from the published co-authorship [12]. The friendship between people is inferred from the information available on their web pages [2]. Discovery of a network structure [18], [16], [17] and detection of an anomaly in a network [20] are also relevant related research topics.

A node discovery predicts the existence of an unknown node around the known nodes from the information on the collective behavior of the network. Related works in the node discovery is limited. Heuristic method for node discovery is proposed in [13]. The method applies clustering algorithm [25], [8] to the nodes in a network, and detects the node which inter-connects clusters at the border of a cluster in clustered networks. The method is applied to analyze the covert social network foundation behind the terrorism disasters [14].

## III. Method

### *A. Observation*

A node and a link in a social network are a person and a relationship resulting in influence transmission between persons. The symbols $n_j$ $(j = 0, 1, \cdots)$ represent the nodes. Some nodes are overt (observable), but the others are covert (unobservable). $\boldsymbol{O}$ denote a set of the whole overt nodes $\{n_0, n_1, \cdots, n_{N-1}\}$. Its cardinality is $N = |\boldsymbol{O}|$. $\boldsymbol{C} = \overline{\boldsymbol{O}}$ denotes a set of the whole covert nodes $\{n_N, n_{N+1}, \cdots\}$. The symbol $\delta_i$ $(0 \leq i < D)$ represent an individual communication

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

pattern (and a resulting collaborative activity) among the persons. It is a set of nodes, $\delta_i \in \boldsymbol{O} \cap \boldsymbol{C}$. The unobservability of the covert nodes does not affect the communication pattern. For example, the members of a communication pattern are those who join an online community.

An observation $d_i$ in surveillance logs is a set of the overt nodes in a communication pattern $\delta_i$. It is given by eq.(1). The number of data is $D$.

$$d_i = \delta_i \cap \boldsymbol{O} \ (0 \leq i < D). \tag{1}$$

$\{d_i\}$ denotes the observation dataset. Note that neither an individual node nor a single link can be observed directly, but a group of nodes can be observed as a communication pattern. $\{d_i\}$ can be expressed by a 2-dimensional $D \times N$ matrix of binary variables $\boldsymbol{d}$. The presence or absence of the node $n_j$ in the data $d_i$ is indicated by the elements in eq.(2).

$$\boldsymbol{d}_{ij} = \begin{cases} 1 & \text{if } n_j \in d_i \\ 0 & \text{otherwise} \end{cases} \ (0 \leq i < D, \ 0 \leq j < N). \tag{2}$$

### *B. Maximal Likelihood Estimator Network*

A parametric form is defined to describe the network topology and the influence transmission over the network. The influence transmission governs the possible communication patterns $\{\delta_i\}$ which result in the observation dataset $\{d_i\}$. The probability where the influence transmits from an initiating node $n_j$ to a responder node $n_k$ is $r_{jk}$. The influence transmits to multiple responders independently in parallel. It is similar to the degree of collaboration probability in trust modeling [11]. The constraints are $0 \leq r_{jk}$ and $\sum_{k \neq j} r_{jk} \leq 1$. The quantity $f_j$ is the probability where the node $n_j$ becomes an initiator. The constraints are $0 \leq f_j$ and $\sum_{j=0}^{N-1} f_j = 1$. These parameters are defined for the whole nodes in a social network (both the nodes in $\boldsymbol{O}$ and $\boldsymbol{C}$).

A single symbol $\boldsymbol{\theta}$ represent both of the parameters $r_{jk}$ and $f_j$ for the nodes in $\boldsymbol{O}$. $\boldsymbol{\theta}$ is the target variable, the value of which needs to be inferred from the observation dataset. The logarithmic likelihood function [8] is defined by eq.(3). The quantity $p(\{d_i\}|\boldsymbol{\theta})$ denote the probability where the observation dataset $\{d_i\}$ realizes under a given $\boldsymbol{\theta}$.

$$L(\boldsymbol{\theta}) = \log(p(\{d_i\}|\boldsymbol{\theta})). \tag{3}$$

The individual observations are assumed to be independent. eq.(3) becomes eq.(4).

$$L(\boldsymbol{\theta}) = \log(\prod_{i=0}^{D-1} p(d_i|\boldsymbol{\theta})) = \sum_{i=0}^{D-1} \log(p(d_i|\boldsymbol{\theta})). \tag{4}$$

The quantity $q_{i|jk}$ in eq.(5) is the probability where the presence or absence of the node $n_k$ as a responder to the stimulating node $n_j$ coincides with the observation $d_i$.

$$q_{i|jk} = \begin{cases} r_{jk} & \text{if } \boldsymbol{d}_{ik} = 1 \text{ for given } i \text{ and } j \\ 1 - r_{jk} & \text{otherwise} \end{cases} . \tag{5}$$

eq.(5) is equivalent to eq.(6) since the value of $d_{ik}$ is either 0 or 1.

$$q_{i|jk} = \boldsymbol{d}_{ik} r_{jk} + (1 - \boldsymbol{d}_{ik})(1 - r_{jk}). \tag{6}$$

The probability $p(\{d_i\}|\boldsymbol{\theta})$ in eq.(4) is expressed by eq.(7). The operator $\wedge$ means logical AND.

$$p(d_i|\boldsymbol{\theta}) = \sum_{j=0}^{N-1} \boldsymbol{d}_{ij} f_j \prod_{0 \leq k < N \ \wedge \ k \neq j} q_{i|jk}. \tag{7}$$

The logarithmic likelihood function takes an explicit formula in eq.(8). The case $k = j$ in multiplication ($\prod_k$) is included since $d_{ik}^2 = d_{ik}$ always holds.

$$L(\boldsymbol{\theta}) = \sum_{i=0}^{D-1} \log(\sum_{j=0}^{N-1} \boldsymbol{d}_{ij} f_j \prod_{k=0}^{N-1} \{1 - \boldsymbol{d}_{ik} + (2\boldsymbol{d}_{ik} - 1) r_{jk}\}). \tag{8}$$

The maximal likelihood estimator $\hat{\boldsymbol{\theta}}$ is obtained by solving eq.(9).

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}). \tag{9}$$

A simple incremental optimization technique (hill climbing method) is employed to solve eq.(9). Simulated annealing method [8] can be employed to strengthen the search ability and to avoid sub-optimal solutions. These methods search more optimal parameter values around the present values and update them as in eq.(10) until the values converge.

$$\begin{cases} r_{jk} \to r_{jk} + \Delta r_{jk} \\ f_j \to f_j + \Delta f_j \end{cases} \ (0 \leq j, k < N). \tag{10}$$

The update $\Delta r_{nm}$ and $\Delta f_n$ should be in the direction of the maximal ascend of the likelihood function. It is indicated by the multiplication of the derivatives and the updates in eq.(11).

$$\Delta L(\boldsymbol{\theta}) = \sum_{n,m=0}^{N-1} \frac{\partial L(\boldsymbol{\theta})}{\partial r_{nm}} \Delta r_{nm} + \sum_{n=0}^{N-1} \frac{\partial L(\boldsymbol{\theta})}{\partial f_n} \Delta f_n. \tag{11}$$

Individual derivatives in eq.(11) are calculated by eq.(12), and eq.(13).

$$\frac{\partial L(\boldsymbol{\theta})}{\partial r_{nm}} = \sum_{i=0}^{D-1} [f_n \boldsymbol{d}_{in} (2\boldsymbol{d}_{im} - 1) \prod_{k \neq m} \{1 - d_{ik} + (2\boldsymbol{d}_{ik} -1) r_{nk}\} \div \sum_{j=0}^{N-1} \boldsymbol{d}_{ij} f_j \prod_{k=0}^{N-1} \{1 - \boldsymbol{d}_{ik} + (2\boldsymbol{d}_{ik} - 1) r_{jk}\}]. \tag{12}$$

$$\frac{\partial L(\boldsymbol{\theta})}{\partial f_n} = \sum_{i=0}^{D-1} [\boldsymbol{d}_{in} \prod_{k=0}^{N-1} \{1 - \boldsymbol{d}_{ik} + (2\boldsymbol{d}_{ik} - 1) r_{nk}\} \div \sum_{j=0}^{N-1} \boldsymbol{d}_{ij} f_j \prod_{k=0}^{N-1} \{1 - \boldsymbol{d}_{ik} + (2\boldsymbol{d}_{ik} - 1) r_{jk}\}]. \tag{13}$$

### *C. Node Discovery - Anomaly Detection*

Suspiciousness of the observation data $d_i$ is evaluated by eq.(14). Suspiciousness means the likeliness where the covert node would appear in the data if it became overt. Larger value means more suspicious data.

$$s(d_i) = \frac{1}{p(d_i|\hat{\boldsymbol{\theta}})}. \tag{14}$$

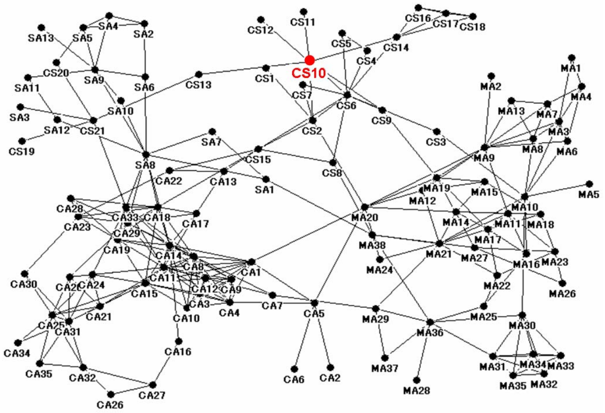

Fig. 1. Network model (A) representing global mujahedin (Jihad fighters) organization [19]. The model consists of 107 nodes, and 4 regional sub-networks. The sub-networks represent Central Staffs (CS), Core Arabs (CA), Maghreb Arabs (MA), and Southeast Asians (SA). The node $n_{\mathrm{CS10}}$, which is indicated by a red circle, is believed to be the founder of the organization.

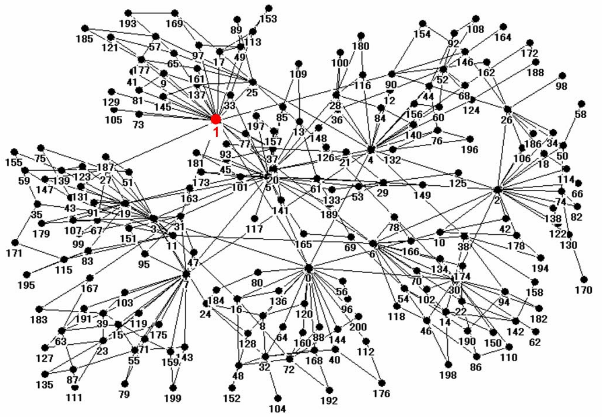

Fig. 2. Network model (C) consisting of 201 nodes and $G = 8$ groups. The group contrast parameter in eq.(16) is $\eta(G-1) = 400$. The node $n_1$, which is indicated by a red circle, is the largest hub whose nodal degree is $K(n_1) = 23$.

Ranking of the observation data can be calculated from the value of eq.(14). The $i$-th most suspicious data is given by $d_{\sigma(i)}$ in eq.(15). Suspiciousness $s(d_{\sigma(i)})$ is larger than $s(d_{\sigma(i')})$ for any $i < i'$.

$$\sigma(i) = \arg \max_{m \neq \sigma(n) \text{ for } \forall n < i} s(d_m) \ (1 \leq i \leq D). \tag{15}$$

## IV. Test Dataset

### A. Network Model

Two classes of network models are employed to generate communication test dataset. The first class is a real organization. The second class is a mathematical model having several adjustable parameters.

The network model (A) in Fig.1 represents a real organization. It is a global mujahedin (Jihad fighters) organization which was analyzed in [19]. The model consists of 107 persons and 4 regional sub-networks. The sub-networks represent Central Staffs (CS), Core Arabs (CA) from the Arabian Peninsula countries and Egypt, Maghreb Arabs (MA) from the North African countries, and Southeast Asians (SA). The organization has a relatively large Gini coefficient of the nodal degree, 0.35, and a relatively large average clustering coefficient, 0.54, [24]. In economics, the Gini coefficient is a measure of inequality of income distribution or of wealth distribution. A larger Gini coefficient indicates lower equality. The values mean that the organization possesses hubs and a group structure.

The node $n_{\mathrm{CS10}}$ (indicated by a red circle in Fig.1) is a hub having relatively large nodal degree ($K(n_{\mathrm{CS10}}) = 8$). It is believed to be the founder of the organization, and said to be the covert leader who provides operational commanders in regional sub-networks with financial support in many terrorism attacks including 9/11 in 2001. His whereabouts are not known despite many efforts in investigation and capture.

The model (A) provides with the practical implication of solving the node discovery problem. But mathematical model is more suitable than the real organization to study the extensive quantitative characteristics of the method. Here, Barabási-Albert model [3] with a group structure is used as a generalization of hubs and group structure in the model (A). The Barabási-Albert model grows with preferential attachment. The probability where a newly coming node $n_k$ connects a link to an existing node $n_j$ is proportional to the nodal degree of $n_j$ ($p(k \to j) \propto K(n_j)$). The occurrence frequency of the nodal degree tends to be scale-free ($f(K) \propto K^a$). In the Barabási-Albert model with a group structure, every node $n_j$ is assigned a pre-determined group attribute $c(n_j)$ to which it belongs. The number of groups is $G$. The probability $p(k \to j)$ is modified to eq.(16). Group contrast parameter $\eta$ is introduced. Links between the groups appear less frequently as $\eta$ increases. The initial links between the groups are connected at random before growth by preferential attachment starts.

$$p(k \to j) \propto \left\{ \begin{array}{ll} \eta(G-1)K(n_j) & \text{if } c(n_j) = c(n_k) \\ K(n_j) & \text{otherwise} \end{array} \right. . \tag{16}$$

The network models (B), (C), and (D) are examples where the number of nodes is 201 and $G = 1$, 8, and 201. The model (B) results in the conventional Barabási-Albert model. The model (C) is shown in Fig.2. The model (D) results in Erdös-Rényi model [6] which is configured by random connection between nodes.

### B. Communication Model

The dataset for performance tests is generated from the network models in IV-A in the 2 steps below.

In the first step, the communication patterns $\{\delta_i\}$ are generated $D$ times according to the influence transmission under the true value of $\boldsymbol{\theta}$. A pattern includes both an initiator node $n_j$ and multiple responder nodes $n_k$. An example is $\delta_0 =$

$\{n_{\mathrm{CS1}}, n_{\mathrm{CS2}}, n_{\mathrm{CS4}}, n_{\mathrm{CS5}}, n_{\mathrm{CS6}}, n_{\mathrm{CS7}}, n_{\mathrm{CS10}}, n_{\mathrm{CS11}}, n_{\mathrm{CS12}}\}$ for the model (A) in Fig.1.

In the second step, the observation dataset $\{d_i\}$ is generated by deleting the covert nodes in $\boldsymbol{C}$ from the patterns $\{\delta_i\}$. The example $\delta_0$ results in the observation $d_0 = \delta_0 \cap \overline{\boldsymbol{C}} = \{n_{\mathrm{CS1}}, n_{\mathrm{CS2}}, n_{\mathrm{CS4}}, n_{\mathrm{CS5}}, n_{\mathrm{CS6}}, n_{\mathrm{CS7}}, n_{\mathrm{CS11}}, n_{\mathrm{CS12}}\}$ if the experimental condition is that $\boldsymbol{C} = \{n_{\mathrm{CS10}}\}$.

The covert node in $\boldsymbol{C}$ may appear multiple times in the communication patterns $\{\delta_i\}$. The number of the target observation data to identify is given by $D_{\mathrm{t}} = \sum_{i=0}^{D-1} B(d_i \neq \delta_i)$. The function $B(s)$ returns 1 if the statement $s$ is true and 0 otherwise. A few conditions are assumed in the performance evaluation in V for simplicity. At first, the probability $f_j$ does not depend on the nodes ($f_j = 1/|\boldsymbol{O} \cup \boldsymbol{C}|$). Second, the value of the probability $r_{ij}$ is either 0 or 1. The number of the possible communication patterns is bounded (less than or equal to the number of nodes $N$). Finally, the influence transmission is bi-directional ($r_{jk} = r_{kj}$).

## V. Performance Evaluation

### A. Performance Measure

Precision, recall, and F measure are used as a measure of the performance. In information retrieval (such as search, document classification, and query classification), the precision $p$ is used as evaluation criteria, which is the fraction of the number of relevant data to the number of the all data retrieved by search. The recall $r$ is the fraction of the number of the data retrieved by search to the number of the all relevant data. The relevant data refers to the data where $d_i \neq \delta_i$. They are given by eq.(17) and eq.(18) They are functions of the number of the retrieved data $D_{\mathrm{r}}$. It can take the value from 1 to $D$. The data is retrieved in the order of $d_{\sigma(1)}$, $d_{\sigma(2)}$, to $d_{\sigma(D_{\mathrm{r}})}$.

$$p(D_{\mathrm{r}}) = \frac{\sum_{i=1}^{D_{\mathrm{r}}} B(d_{\sigma(i)} \neq \delta_{\sigma(i)})}{D_{\mathrm{r}}}. \tag{17}$$

$$r(D_{\mathrm{r}}) = \frac{\sum_{i=1}^{D_{\mathrm{r}}} B(d_{\sigma(i)} \neq \delta_{\sigma(i)})}{D_{\mathrm{t}}}. \tag{18}$$

The F measure $F$ is the harmonic mean of the precision and recall [10]. It is given by eq.(19).

$$F(D_{\mathrm{r}}) = \frac{1}{\frac{1}{2}(\frac{1}{p(D_{\mathrm{r}})} + \frac{1}{r(D_{\mathrm{r}})})} = \frac{2p(D_{\mathrm{r}})r(D_{\mathrm{r}})}{p(D_{\mathrm{r}}) + r(D_{\mathrm{r}})}. \tag{19}$$

The precision, recall, and F measure range from 0 to 1. All the measures take larger values as the performance of retrieval becomes better.

### B. Result

The results of the performance evaluation using the test dataset in IV-B derived from the network models in IV-A are demonstrated.

Let's start with the first class of the network models (real organization) and learn the implication of the method. Fig.3 shows the precision ($p$), recall ($r$), and F measure ($F$) in the trial where the experimental condition is that the node $n_{\mathrm{CS10}}$ in the model (A) is the target covert node to discover ($\boldsymbol{C} = \{n_{\mathrm{CS10}}\}$, $|\boldsymbol{C}| = 1$, $N = |\boldsymbol{O}| = 106$). The horizontal axis is the rate of the number of the retrieved data ($D_{\mathrm{r}}$) to the number of the whole data ($D$). The vertical solid line indicates the position at $D_{\mathrm{r}} = D_{\mathrm{t}}$. The broken lines indicate the theoretical limit (upper bound) and the random retrieval limit (lower bound). The evaluation is under the condition where the all possible communication patterns are known.

The precision, recall, and F measure are the same value of 0.78 at $D_{\mathrm{r}} = D_{\mathrm{t}}$. These are much better than those of the random retrieval ($F(D_{\mathrm{t}}) = 0.04$) and close to the theoretical limit. The method fails to discover two suspicious records $\delta_i$ =$\{n_{\mathrm{CS10}}$, $n_{\mathrm{CS11}}\}$, and $\{n_{\mathrm{CS10}}$, $n_{\mathrm{CS12}}\}$ when $D_{\mathrm{r}}$ is small. This indicates that the communication with the nodes having small nodal degree ($K(n_{\mathrm{CS11}}) = 1$ and $K(n_{\mathrm{CS12}}) = 1$) does not provide much clues for node discovery. On the other hand, the most suspicious observation data $d_{\sigma(1)}$ includes all the neighbor nodes $n_{\mathrm{CS1}}$, $n_{\mathrm{CS2}}$, $n_{\mathrm{CS4}}$, $n_{\mathrm{CS5}}$, $n_{\mathrm{CS6}}$, $n_{\mathrm{CS7}}$, $n_{\mathrm{CS11}}$, and $n_{\mathrm{CS12}}$. The method succeeded in discovering most of the suspicious records and the all suspicious nodes. The investigators will decide to collect more detailed information on the communication (and a resulting collaborative activity) of the suspicious neighbor nodes. This will result in identifying, locating, and finally, capturing the covert leader ($\boldsymbol{C} = \{n_{\mathrm{CS10}}\}$) who is responsible for many terrorism attacks.

Let's move on to the second class of the network model (mathematical model with adjustable parameters) and study the extensive performance characteristics of the method. Fig.4 shows the F measure $F(D_{\mathrm{t}})$ as a function of the nodal degree $K$. Individual plots shows the F measure averaged over the trials where the experimental condition is that a node having a given nodal degree $K(n_i) = K$ is the target covert node to discover ($N = |\boldsymbol{O}| = 200$, $|\boldsymbol{C}| = 1$). The solid line graphs (a), (b), and (c) are for the model (B), (C), and (D). The broken lines indicate the theoretical limit and the random retrieval limit. The evaluation is under the condition where the all possible communication patterns are known in Fig.4 through Fig.6. The resulting F measure ranges from 0.7 to 1. It does not depend on the number of groups (or topology of the network model). The performance becomes better as the nodal degree of the target covert nodes increases.

Fig.5 shows the F measure $F(D_{\mathrm{t}})$ as a function of the number of groups $G$ in the trial where the largest hub is the target covert node to discover ($N = |\boldsymbol{O}| = 200$, $|\boldsymbol{C}| = 1$). The horizontal axis is $G/N$. The number of the nodes is constant. The group contrast parameter is fixed at $\eta(G-1) = 400$ regardless of the value of $G$. The broken lines indicate the theoretical limit and the random retrieval limit. The F measure degrades down to 0.7 around $G = 0.5N$. But the performance still remains much better than that of the random retrieval. The method can be applied for any value of $G$.

Fig.6 shows the F measure $F(D_{\mathrm{t}})$ as a function of the number of overt nodes ($N = |\boldsymbol{O}|$) in the trial where the largest hub is the target covert node to discover ($|\boldsymbol{C}| = 1$). The number of the groups is constant ($G = 1$). The broken lines indicate the theoretical limit and the random retrieval limit. Except the case $N = 50$, the performance remains

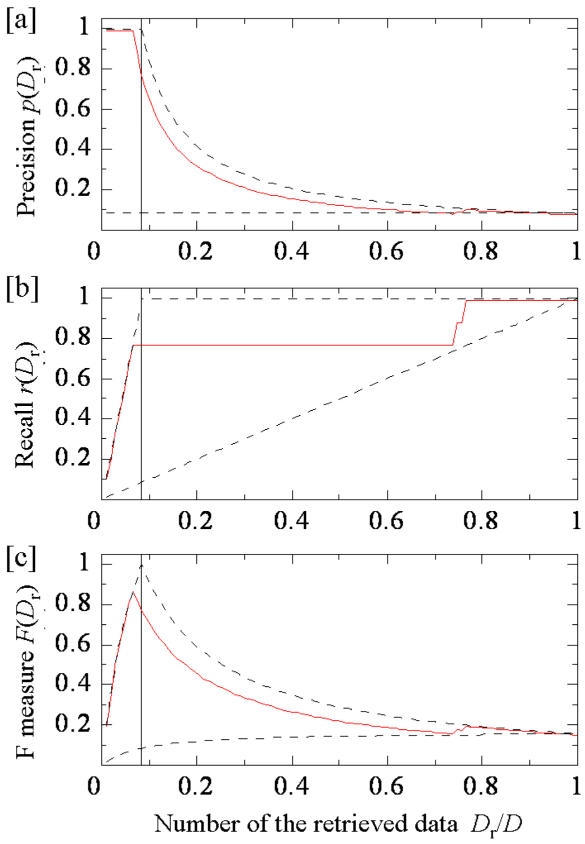

Fig. 3. Precision ($p$), recall ($r$), and F measure ($F$) in the trial where the node $n_{\mathrm{CS10}}$ in the model (A) is the target covert node to discover ($\boldsymbol{C} = \{n_{\mathrm{CS10}}\}$, $|\boldsymbol{C}| = 1$, $N = |\boldsymbol{O}| = 106$). The horizontal axis is the rate of the number of the retrieved data ($D_{\mathrm{r}}$) to the number of the whole data ($D$). The vertical solid line indicates the position at $D_{\mathrm{r}} = D_{\mathrm{t}}$. The broken lines indicate the theoretical limit and the random retrieval limit.

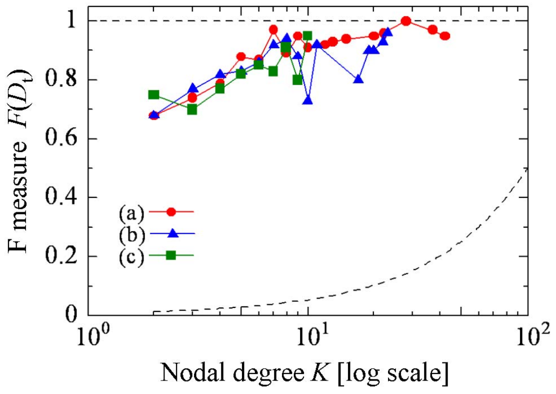

Fig. 4. F measure $F(D_{\mathrm{t}})$ as a function of the nodal degree $K$. Individual plots shows the F measure averaged over the trials where a node having a given nodal degree $K(n_i) = K$ is the target covert node to discover ($|\boldsymbol{C}| = 1$). The solid line graphs (a), (b), and (c) are for the model (B), (C), and (D) ($|\boldsymbol{O}| = 200$). The broken lines indicate the theoretical limit and the random retrieval limit.

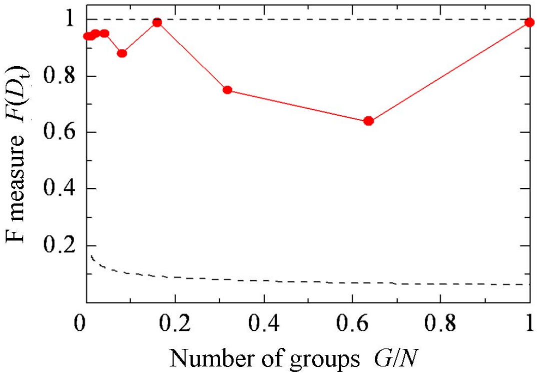

Fig. 5. F measure $F(D_{\mathrm{t}})$ as a function of the number of groups $G$ in the trial where the largest hub is the target covert node to discover. The number of the nodes is constant ($|\boldsymbol{O}| = 200$, $|\boldsymbol{C}| = 1$). The group contrast parameter is $\eta(G-1) = 400$. The broken lines indicate the theoretical limit and the random retrieval limit.

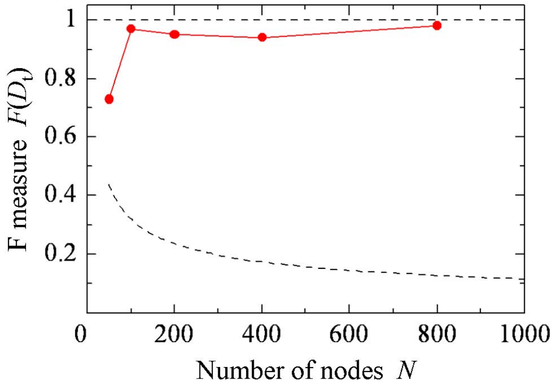

Fig. 6. F measure $F(D_{\mathrm{t}})$ as a function of the number of nodes $N = |\boldsymbol{O}|$ in the trial where the largest hub is the target covert node to discover ($|\boldsymbol{C}| = 1$). The number of the groups is constant ($G = 1$). The broken lines indicate the theoretical limit and the random retrieval limit.

constant. The method can be applied for large networks. Note that several hours of calculation is necessary with a standard personal computer when $N$ approaches to 1000. The size of the network is limited by the amount of calculation rather than by the accuracy obtainable from the method.

Fig.7 shows the F measure $F(D_{\mathrm{t}})$ as a function of the number of the observed data $D$ in the trial where the node $n_0$ in the model (B) is the target covert node to discover ($\boldsymbol{C} = \{n_0\}$). The horizontal axis is the ratio of $D$ to the number of the possible communication patterns ($|\boldsymbol{O} \cup \boldsymbol{C}| = N+1$) as assumed in IV-B. The ratio was $D/N = 1$ in Fig.4 through Fig.6. The broken lines indicate the theoretical limit and the random retrieval limit. The F measure is close to the theoretical limit, if more than 80% of the possible communication patterns is observed. The performance is no better than that of the random retrieval if only 50% of the possible communication patterns is observed. It is a major restriction imposed on the method that many of the possible communication patterns need to be

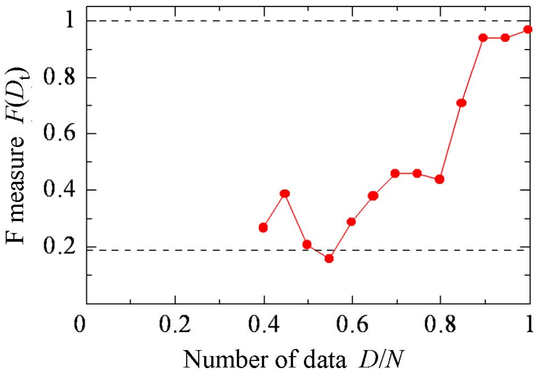

Fig. 7. F measure $F(D_\mathrm{t})$ as a function of the number of the observed data $D$ in the trial where the node $n_0$ in the model (B) is the target covert node to discover. The horizontal axis is $D/N$. The broken lines indicate the theoretical limit and the random retrieval limit.

known. Overcoming the restriction is for future work.

## VI. Conclusion

In this paper, I define a node discovery problem in a networked organization and present a method to solve the problem. The method infers the network behind the social interactions, applies an anomaly detection technique to the surveillance logs, and identifies the suspicious surveillance logs. The precision, recall, and F measure characteristics are close to the theoretical limit for any covert nodes in the networks of any topologies and sizes. I believe that, in the investigation of a clandestine organization [14], the method aids the investigators in identifying the close associates (participants in the most suspicious surveillance record) of a covert leader or a critical conspirator.

I plan to address 3 issues for the future works. The first issue is to overcome the restriction where the performance degrades unless the ratio of the number of the observation to the number of possible communication patterns is large. The second issue is to extend the models for the social interactions. The model in this paper represents the radial influence transmission from an initiating node toward multiple responder nodes. In real networked organizations, other types of influence transmission are present such as serial (chain-shaped) influence transmission, or tree-like influence transmission. The third issue is to develop a method to solve the variants of the node discovery problem. Discovering fake nodes, or spoofing nodes are also interesting problems to uncover the malicious intentions of the organization. A fake node is the person who does not exist in the organization, but appears in the surveillance. A spoofing node is the person who belong to an organization, but appears as a different node in the surveillance.

We encounter the node discovery problem in many areas of business and social sciences [15]. For example, in document analysis [23], something unknown, which is not stated explicitly, can be discovered. The discovery may provide the analyst with a clue to approach the hidden intention of the author, an opinion which is about to emerge, or a sign of trends. The method will be the new basis to analyze something hidden behind the direct observation, which is beyond the scope of the conventional statistical methods and data mining expertises.

## References

[1] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. Huberman, Search in power-law networks, Physical Review E vol. 64, 046135, 2001.

[2] L. A. Adamic, and E. Adar, Friends and neighbors on the web, *Social Networks* vol. 25, pp. 211-228, 2003.

[3] A. L. Barabási, R. Albert, and H. Jeong, Mean-field theory for scale-free random networks, *Physica A* vol. 272, pp. 173-187, 1999.

[4] M. Chau, and H. Chen, Incorporating web analysis into neural networks: an example in hopfield net searching, *IEEE Transactions on Systems, Man, & Cybernetics Part C* vol. 37, pp. 352-358, 2007.

[5] A. Clauset, C. Moore, and M. E. J. Newman: Hierarchical structure and the prediction of missing links in networks, *Nature* vol. 453, pp. 98-100, 2008.

[6] P. Erdös, and A. Rény, On random graphs. I., *Publicationes Mathematicae* vol. 6, pp. 290-297, 1959.

[7] L. Getoor, and C. P. Diehl, Link mining: a survey, *ACM SIGKDD Explorations* vol. 7, pp. 3-12, 2005.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning - Data mining, inference, and prediction*. Springer-Verlag, 2001.

[9] P. S. Keila, and D. B. Skillicorn, Structure in the Enron email dataset, *Journal of Computational & Mathematical Organization Theory* vol. 11, pp. 183-199, 2006.

[10] R. R. Korfhuge, *Information storage and retrieval*. Wiley, 1997.

[11] N. Lavrač, P. Ljubič, T. Urbančič, G. Papa, M. Jermol, and S. Bollhalter, Trust modeling for networked organizations using reputation and collaboration estimates, *IEEE Transactions on Systems, Man, & Cybernetics Part C* vol. 37, pp. 429-439, 2007.

[12] D. Liben-Nowell, and J. Kleinberg, The link prediction problem for social networks, *Journal of American Society of Information Science and Technology* vol. 58, pp.1019-1031, 2007.

[13] Y. Maeno, and Y. Ohsawa, Human-computer interactive annealing for discovering invisible dark events, *IEEE Transactions on Industrial Electronics* vol. 54, pp. 1184-1192, 2007.

[14] Y. Maeno, and Y. Ohsawa, Analyzing covert social network foundation behind terrorism disaster, *International Journal of Services Sciences* vol. 2, pp. 125-141, 2009.

[15] Y. Maeno, and Y. Ohsawa, Reflective visualization and verbalization of unconscious preference, in press, *International Journal of Advanced Intelligence Paradigms*, 2009. Available e-print http://arxiv.org/abs/0803.4074.

[16] M. E. J. Newman, and E. A. Leicht: Mixture models and exploratory analysis in networks, *Proceedings of the National Academy of Sciences USA* vol. 104, pp. 9564-9569, 2007.

[17] G. Palla, I. Derènyi, I Farkas, and T. Vicsek: Uncovering the overlapping community structure of complex networks in nature and society, *Nature* vol. 435, pp. 814-818, 2005.

[18] M. G. Rabbat, M. A. T. Figueiredo, and R. D. Nowak: Network Inference from co-occurrences, *IEEE Transactions on Information Theory* vol. 54, pp. 4053-4068, 2008.

[19] M. Sageman, *Understanding terror networks*. University of Pennsylvania Press, 2004.

[20] J. Silva, and R. Willett: Hypergraph-based anomaly detection of high-dimensional co-occurrences, *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. 31, pp. 563-569, 2009.

[21] R. Silva, R. Scheines, C. Glymour, and P. Spirtes, Learning the structure of linear latent variable models, *Journal of Machine Learning Research* vol. 7, pp. 191-246, 2006.

[22] S. Singh, J. Allanach, T. Haiying, K. Pattipati, and P. Willett, Stochastic modeling of a terrorist event via the ASAM system, in *Proceedings of the IEEE International Conference on Systems, Man & Cybernetics, Hague*, 2004.

[23] B. Taskar, M. F. Wong, P. Abbeel, and D. Koller, Link prediction in relational data, in *Proceedings of the Neural Information Processing Systems Conference, Vancouver*, 2003.

[24] D. J. Watts, and S. H. Strogatz, Collective dynamics of small-world networks, *Nature* vol. 398, pp. 440-442, 1998.

[25] A. Zakarian, A new nonbinary matrix clustering algorithm for development of system architectures, *IEEE Transactions on Systems, Man, & Cybernetics Part C* vol. 38, pp. 135-141, 2008.