Generalizing Terrorist Social Networks with K-Nearest Neighbor and Edge Betweeness for Social Network Integration and Privacy Preservation

Xuning Tang
College of Information Science and Technology
Drexel University
3141 Chestnut Street
Philadelphia, PA 19104
xt24@drexel.edu

Christopher C. Yang
College of Information Science and Technology
Drexel University
3141 Chestnut Street
Philadelphia, PA 19104
chris.yang@drexel.edu

Abstract—Social network analysis has been shown to be effective in supporting intelligence and law enforcement force to identify suspects, terrorist or criminal subgroups, and their communication patterns. However, social network data owned by individual law enforcement units contain private information that must be preserved before sharing with other law enforcement units. Such privacy issue tremendously reduces the utility of the social network data since the integration of social networks from different law enforcement units cannot be fully integrated. Without integration of social network data, the effectiveness of terrorist or criminal social network analysis is diminished. In this paper, we introduce the *KNN* and *EBB* algorithm for constructing generalized subgraphs and a mechanism to integrate the generalized information to conduct the closeness centrality measures. The result shows that the proposed technique improves the accuracy of closeness centrality measures substantially while protecting the sensitive data.

1. INTRODUCTION

As social network data is becoming more publicly available due to the advance of online social networking, social network analysis techniques are attracting increasing research interest in both academia and industry. In intelligence and security informatics domain, social network analysis techniques have been widely used to support intelligence and law enforcement force to identify suspects, gateways, and extracting communication patterns of terrorist or criminal organizations. In our previous work [21] [23], we have shown how social network analysis and visualization techniques are useful in knowledge discovery of terrorist social network.

Social network analysis is useful for extracting the complex relationships between social actors. However, in practice, this capability is greatly compromised to deal with the limitation of partial data or anonymized data. It is understandable that the social network analysis techniques are not able to extract the essential knowledge by using partial data of a terrorist or criminal social network. For example, each law enforcement unit has its own criminal social network collected by its own agencies but this network is only part of the global criminal social network. Mining on an incomplete criminal social network may not be able to identify the bridge between two criminal subgroups. Ideally, information sharing can resolve the problem. However, due to the privacy policy, different organizations are not allowed to share the sensitive information of their social network data. As a result, they have to anonymize the social network data before publishing or sharing it. But, even the modest privacy gains require almost complete destruction of the data-mining utility, which means it is hard to make a balance between privacy and utility in network data publishing.

An accurate social network analysis cannot be conducted unless an integration of the social networks owned by different organizations can be made. To illustrate this idea, assuming organization P (O_P) and organization Q (O_Q) own the social networks G_P and G_Q respectively as shown in Figure 1, without integrating G_P and G_Q, O_P will never discover the close relationship between the A and G. Similarly, both O_P and O_Q will never discover the connection between C and D. After integrating G_P and G_Q to $\boldsymbol{G}$, both O_P and O_Q will identify the connections.

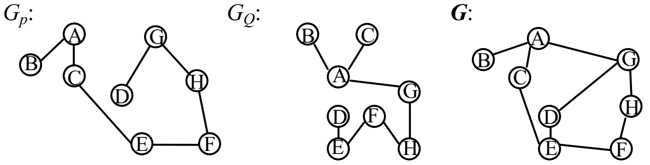

Figure 1. Illustrations of social networks integration

In this paper, the research problem is defined as follow. Given two or more social networks (G_1, G_2, …) from different organizations (O_1, O_2, …), the objective is *sharing the necessary information* between these social networks to achieve a more *accurate social network analysis and mining result* and *preserving the sensitive information* at the same time. Each organization O_i has a piece of social network, which is part of the whole picture – a social network $\boldsymbol{G}$ constructed by integrating all G_i. Conducting the social network analysis task on $\boldsymbol{G}$, one can obtain the exact result. However, conducting the social network analysis task on any G_i, one can never achieve the exact social network analysis result because of the missing information. By integrating G_i and the generalized information of G_j, O_i should be able to achieve a more accurate SNAM result.

In this paper, we propose the algorithms for social network data sharing and integration. The proposed information sharing and integration of social networks have three major components, (i) constructing generalized subgraph, (ii) creating generalized information for sharing, and (iii) social networks integration and analysis. In particular, we propose and investigate two subgraph generalization algorithms, edge betweenness based (EBB) and K-nearest neighbor (KNN). We evaluated the effectiveness of these algorithms using the Global Salafi Jihad terrorist social network. The experimental result shows that subgraph generalization is an effective and pragmatic way for information integration while preserving the data privacy.

1.1 Related Work

Assured has been studied for many years [2][3][16][17]. Baird et al. [2],[3] first discussed several aspects of coalition data sharing in the Markle report. Thuraisingham [16][17] defined assured information sharing as enforcing security and integrity policies during information sharing between organizations so that the data is integrated and mined to extract nuggets. Thuraisingham [16] has further discussed these

978-1-4244-6446-3/10/$26.00 © 2010 IEEE

aspects including confidentiality, privacy, trust, integrity, dissemination and others. In this work, we focus on social network data sharing and integration. Sensitive information should be protected while insensitive information can be shared or used to generate generalized information so that privacy can be protected and generalized information can be integrated for social network analysis.

A number of anonymity approaches for preserving privacy of relational data have been investigated in the recent years. These techniques include k-anonymity [12],[14], l-diversity [9], t-closeness [7], m-invariance [20], and δ-presence [10]. *k-anonymity* [12],[14] ensures at least k records with respect to every set of quasi-identifier attributes are indistinguishable. However, k-anonymity fails when there is a lack of diversity or other background knowledge. *l-diversity* [9] ensures that there are at least l well-represented values of the attributes for every set of quasi-identifier attributes. The weakness is that one can still estimate the probability of a particular sensitive value. *m-invariance* [20] ensures that each set of quasi-identifier attributes has at least m tuples, each with a unique set of sensitive values. Personalization is added other enhanced techniques such as personalized anonymity [19] and (α,k)-anonymity [18]. These techniques allow users to specify the degree of privacy protection or specify a threshold α on the relative frequency of the sensitive data. In this work, we focus on the privacy preservation of social network data rather than relational data. The technique in privacy preservation of relational data is not directly applicable in social network data because the data representations are different.

Relatively less research work on privacy preservation of social network data (or graphs) has been done. A naïve approach is removing the identities of all nodes but only revealing the edges of a social network. There are several anonymization models proposed in the recent literature: *k-candidate anonymity* [7], *k-degree anonymity* [9], *and k-anonymity* [25]. Such anonymization models are proposed to increase the difficulty of being attacked based on the notion of k-anonymity in tabular data. k-candidate anonymity [7] defines that there are at least k candidates in a graph G that satisfies a given query Q. k-degree anonymity [9] defines that, for every node v in a graph G, there are at least k-1 other nodes in G that have the same degree as v. k-anonymity [25] has the strictest constraint. It defines that, for every node v in a graph G, there are at least k-1 other nodes in G such that their anonymized neighborhoods are isomorphic. The technique to achieve the above anonymities is *edge or node perturbation* in addition to removing all identities [7],[9],[25]. By adding and/or deleting edges and/or nodes, a perturbed graph is generated to satisfy the anonymity requirement. Adversaries can only have a confident of $1/k$ to discover the identity of a node by neighborhood attacks.

Using the anonymization approach, the global network properties are mostly preserved for other research applications assuming that the identities of nodes are not of interest in the research applications. By using this assumption, one can study the network property of each anonymized social network separately but we cannot integrate the social networks. On the other hand, Backstorm et al. [1] proved that it is possible to discover whether edges between specific targeted pairs of nodes exist or not by active or passive attacks. Based on the uniqueness of small random subgraphs embedded in a social network, one can infer the identities of nodes by solving a set of restricted isomorphism problems. In addition, the sets of nodes and edges in a perturbed social network are different from the sets of nodes and edges in the original social network. As reported by Zhou and Pei [25], the number of edges added can be as high as 6% of the original number of edges in a social network. A recent study [24] has investigated how edge and node perturbation can change certain network properties. Such distortion may cause significant errors in certain social network analysis tasks such as centrality measurement although the global properties can be maintained.

2. A Framework for Integration Social Networks with Privacy Preservation

Assuming organization P (O_P) has a social network G_P and organization Q (O_Q) has a social network G_Q, O_P needs to conduct a social network analysis (SNA) task but G_P is only a partial social network for the SNA task. If there is not any privacy concern, one can integrate G_P and G_Q to generate an integrated $\boldsymbol{G}$ and obtain a better SNA result. Due to privacy concern, O_Q cannot release G_Q to O_P but only shares some data to O_P according to the agreed privacy policy. At the same time, O_P does not need all data from O_Q but only those that are critical for the Social Network Analysis and Mining (**SNAM**) task. When we integrate social networks with privacy concern, we need to maximize the information sharing that is insensitive and useful for the SNAM tasks. The shared information should not include any sensitive information; however, it must be useful for improving the performance of the SNAM task conducted by the information requesting party.

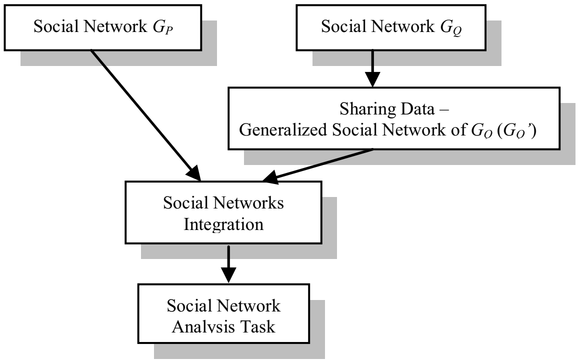
Social Network G_P
Social Network G_Q
Sharing Data –
Generalized Social Network of G_O (G_O')
Social Networks Integration
Social Network Analysis Task

Figure 2. Framework of Social Networks Integration

Using subgraph generalization, a generalized social network, G_Q', will be created from G_Q and conforming to the privacy policy. The generalized social graph only contains generalized information of G_Q without releasing any sensitive information. For instances, the generalized information can be the maximum or minimum length of the shortest paths between two subgroups, the degree of an insensitive node, the radius of a subgroup, etc. The generalized social network G_Q' will then be integrated with G_P to support a social network analysis and mining task. Given the generalized information from G_Q, it is expected to achieve better social network analysis and mining than conducting the analysis and mining on G_P alone.

2.1 Subgraph Generalization

Given the insensitive data in G_Q, we propose and investigate two subgraph generalization methods, K-nearest neighbor (*KNN*) method and Edge betweenness based (EBB) method, to generate a generalized social network G_Q' for sharing with O_P. A subgraph generalization method will create a generalized version of a social network, in which a connected subgraph is transformed as a *generalized node* and only *generalized information* will be presented in the generalized node. A mechanism is needed to (i) identify the subgraphs for generalization, (ii) determine the connectivity between the set of generalized nodes in the generalized social network, and (iii) construct the generalized information to be shared. The generalized social network protects all sensitive information while releasing the crucial and non-sensitive information to the information requesting party for social network integration and the intended SNAM task.

Given a social network, $G = (V, E)$, where V is a set of nodes, E is a set of edges and $|V| = n$, K of these nodes are insensitive nodes. We decompose G into K subgraphs $G_i = (V_i, E_i)$ where each subgraph has one insensitive node. Each subgraph will be transformed as a generalized node in the generalized graph G'. $V = \bigcup_{i=1\text{to}K} V_i$. v_i^C corresponds to the center of a sub-graph G_i. We propose and investigate *KNN* and *EBB* to decompose G into K subgraphs, generalize these subgraphs as generalized nodes, and generate the generalized graph G' by connecting the generalized nodes.

2.1.1 K-nearest neighbor (KNN) method

There are two basic principles in KNN method. Let $SP^D(v, v_i^C)$ be the distance of the shortest path between v and v_i^C. When v is assigned to the subgraph G_i in subgraph generation, $SP^D(v, v_i^C)$ must be shorter than or equal to $SP^D(v, v_j^C)$ where $j = 1, 2, .., K$ and $j \neq i$. Secondly, an edge exists between two generalized nodes G_i and G_j in the generalized graph G' if and only if there is an edge between any two nodes in G such that one from each generalized node, G_i and G_j.

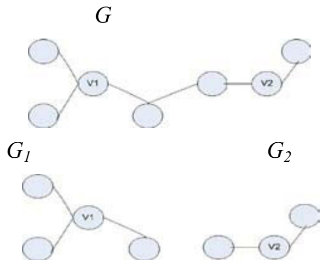

Figure 3. Illustrations of generating subgraphs

For simplicity, we use the graphs in Figure 3 to illustrate the subgraph generation by *KNN* method. G has seven nodes including v_1 and v_2. If we take v_1 and v_2 as the insensitive nodes and we are going to create 2 subgraphs by *1NN* method, all other nodes will be assigned to one of the two subgraphs depending on their shortest distances with v_1 and v_2. Two subgraphs G_1 and G_2 are generated as illustrated in Figure 3. This illustration is made only for simplicity. A real social network will have significantly more number of nodes and edges.

The *KNN* subgraph generation algorithm is presented below:

```
length=1;                                                    Step 1
V= V - {$v_1^C$, $v_2^C$, … $v_K^C$ };                       Step 2
While V ≠ Ø                                                  Step 3
  For each $v_j \in V$                                       Step 4
    For each $i$ = 1 to $K$                                  Step 5
        IF($SP^D(v_j, v_i^C)$ == length);                    Step 6
            $V_i = V_i + v_j$;                               Step 7
            $V = V - v_j$;                                   Step 8
    End For;                                                 Step 9
  End For;                                                   Step 10
length++;                                                    Step 11
End While                                                    Step 12
For each $(v_i,v_j) \in E$                                   Step 13
  IF( Subgraph($v_i$) == Subgraph($v_j$) )                   Step 14
  // Subgraph($v_i$) is the subgraph such that $v_i \in$ Subgraph($v_i$)
    $G_k$ = Subgraph($v_i$)                                  Step 15
    $E_k = E_k + (v_i,v_j)$                                  Step 16
  ELSE                                                       Step 17
    Create an edge between Subgraph($v_i$) and Subgraph($v_j$) and add
it to E'                                             Step 18
End For                                                      Step 19
```

The *KNN* subgraph generation algorithm creates K subgraphs G_1, G_2, …, G_K from G. Each subgraph, G_i, has a set of nodes, V_i, and a set of edges, E_i. Edges between subgraphs, E', are also created. A generalized graph, G', is constructed where each generalized node corresponds a subgraph G_i and labeled by the insensitive node, v_i^C. Using the example in Figure 4, the generalized graph is presented in Figure 4.

Figure 4. Illustrations of a generalized graph

2.1.2 Edge betweenness based (EBB) method

In EBB, we focus on those edges which have the highest betweeness. We construct a generalized graph by progressively removing edges with the highest betweenness from the original graph.

The betweenness of an edge is defined as the number of shortest paths between pairs of nodes that pass through it. Girvan and Newman [6] used edge betweenness to measure the influence of the edge over the flow of information between nodes. If a network contains communities that are only loosely connected by a few numbers of inter-community edges, the shortest paths between different communities are likely to pass through one of these few edges. As a result, the edges that connect communities together usually have a higher edge betweenness. We propose the EBB method based on this principle. EBB decomposes a graph into several subgraphs and ensures that each subgraph contains only one insensitive node.

The EBB algorithm is presented as follows:

```
// EBB(G), Edge Betweenness Based method
Initialize e={};                                              Step 1
While(there are more than one insensitive node in graph G)
    Identify edge ($v_i$,$v_j$) in G which is not an element of e and has
        the highest betweenness;                              Step 2
    Remove ($v_i$,$v_j$) from G;                              Step 3
    IF(G is still connected after removing edges ($v_i$,$v_j$))
      EBB(G);                                                 Step 4
    ELSE IF (G is disconnected and split to two graph $G_p$ and $G_q$)
      IF(No insensitive node in $G_p$) or (No insensitive node in $G_q$)
        Add ($v_i$,$v_j$) back to G;                          Step 5
        e = e+ ($v_i$,$v_j$);                                 Step 6
        Go Back to Step 2;                                    Step 7
      ELSE
        EBB($G_P$);                                           Step 8
        EBB($G_q$);                                           Step 9
End While;

//Add edge between generalized node to form generalized graph
For each $(v_i, v_j) \in E$
    IF( Subgraph($v_i$) == Subgraph($v_j$) )
      $G_k$ = Subgraph($v_i$)
      $E_k = E_k + (v_i,v_j)$
    ELSE
      Create an edge between Subgraph($v_i$) and Subgraph($v_j$) and add
it to E'
End For
```

For example, as shown in Figure 5, the edge connecting two insensitive nodes, v_1 and v_2, has the highest betweenness. At the same time, removing this edge will not generate subgraphs without any insensitive node. EBB removes this edge, creates two subgraphs such that each of them has one insensitive node, generalizes these two subgraphs as two generalized nodes, and creates an edge between these generalized nodes to construct the generalized graph as shown in Figure 5.

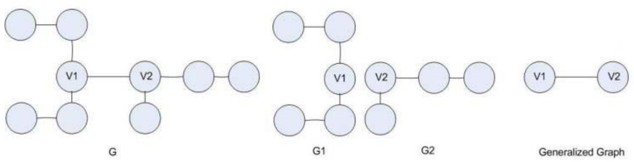

Figure 5. Illustrations of generalizing subgraph byEBB

When an organization is sharing its social network with another organization, the generalized graph is shared but not all information within each generalized node will be shared. Only generalized subgraph information is shared so that sensitive information is preserved. At the same time, generalized information will support integration and social network analysis. In the next section, we describe the generalized subgraph information for sharing.

2.2 Generalized Subgraph Information

Given a generalized node G_i and its center (insensitive node) v_i^C, we propose to create generalized subgraph information such as the shortest paths in a subgraph, the number of nodes in a subgraph and the number of nodes adjacent to another subgraph. The generalized information should not reveal sensitive information including sensitive identities and sensitive relationships. The generalized information, however, should provide useful information for social network analysis after integration.

Let v_p and v_q be any two nodes in G_i and the length of the shortest path between v_p and v_q be $SP^D(v_p,v_q,G_i)$. By considering all the shortest paths between any two nodes in G_i, $SP^D(v_p,v_q,G_i)$ $\forall$ $v_p,v_q \in V_i$, we compute the longest length of the shortest paths between any two nodes in G_i, denoted as $L\_SP^D(G_i)$, and the shortest length of the shortest paths between any two nodes in G_i, denoted as $S\_SP^D(G_i)$.

$$L\_SP^D(G_i) = \{SP^D(v_m,v_n,G_i) \mid \forall\ v_p,v_q \in V_i,\ \ SP^D(v_m,v_n,G_i) \geq SP^D(v_p,v_q,G_i)\}$$

and

$$S\_SP^D(G_i) = \{SP^D(v_m,v_n,G_i) \mid \forall\ v_p,v_q \in V_i,\ \ SP^D(v_m,v_n,G_i) \leq SP^D(v_p,v_q,G_i)\}$$

The length of any shortest paths in G_i, α, must be smaller or equal to $L\_SP^D(G_i)$ and larger or equal to $S\_SP^D(G_i)$.

$$S\_SP^D(G_i) \leq \alpha \leq L\_SP^D(G_i)$$

We can also compute the probability of the length of the shortest path between any two nodes in G_i, denoted as $Prob(SP^D(G_i) = \alpha)$, and $0 \leq Prob(SP^D(G_i)=\alpha) \leq 1$

Similary, let the length of the shortest path between v_p and the center of G_i, v_i^C, be $SP^D(v_p,v_i^C,G_i)$. By considering all the shortest paths between any node in G_i and v_i^C, we compute the longest length and the shortest length of the shortest paths between v_p and v_i^C, denoted as $L\_SP^D(v_i^C,G_i)$ and $S\_SP^D(v_i^C,G_i)$.

$$L\_SP^D(v_i^C,G_i) = \{SP^D(v_m,v_i^C,G_i) \mid\ v_p \in V_i,\ ,\ \ SP^D(v_m,v_i^C,G_i) \geq SP^D(v_p,v_i^C,G_i)\}$$

and

$$S\_SP^D(v_i^C,G_i) = \{SP^D(v_m,v_i^C,G_i) \mid\ v_p \in V_i,\ \ SP^D(v_m,v_i^C,G_i) \leq SP^D(v_p,v_i^C,G_i)\}$$

The probability of the length of the shortest path between any node and v_i^C, denoted as $Prob(SP^D(v_i^C,G_i) = \beta)$, can also be computed. $S\_SP^D(v_i^C, G_i) \leq \beta \leq L\_SP^D(v_i^C, G_i)$ and $0 \leq Prob(SP^D(v_i^C, G_i)= \alpha) \leq 1$,

We denote $Num(G_i)$ as the number of nodes in G_i and $Num(G_i,G_j)$ as the number of nodes in G_i that are adjacent to another subgraph G_j.

The generalized subgraph information for sharing includes: (i) $L\_SP^D(G_i)$, (ii) $S\_SP^D(G_i)$, (iii) $Prob(SP^D(G_i)=\alpha)$, (iv) $L\_SP^D(v_i^C,G_i)$, (v) $S\_SP^D(v_i^C,G_i)$, (vi) $Prob(SP^D(v_i^C,G_i) = \beta)$, (vii) $Num(G_i)$, and (viii) $Num(G_i ,G_j)$.

2.3 Generalized Graph Integration and Social Network Analysis

Given the generalized graph G_Q' and its generalized subgraph information, O_P wants to integrate G_Q' with its own graph G_P to conduct more accurate closeness centrality. In order to achieve such purpose, we need to make estimation of the distance between any two nodes v_i and v_j in G_P by integrating the generalized subgraph information of G_Q'.

To estimate the distance between two nodes v_i and v_j in G_P, we identify the two closest insensitive nodes for v_i and v_j in G_P and use the generalized information from G_Q' to estimate the distances between the insensitive nodes. We only consider the two closest insensitive nodes but the formulation can be easily modified to consider other insensitive nodes that are further away. Let the closest insensitive node to v_i in G_P be v_A^c, and the second closest insensitive node to v_i in G_P be $v_{A'}^c$. We set the weights λ_A and $\lambda_{A'}$ as

$$\lambda_A = \frac{SP^D\left(v_i ,v_{A'}^c ,G_P\right)}{SP^D(v_i ,v_A^c ,G_P)+SP^D\left(v_i ,v_{A'}^c ,G_P\right)} \text{ and}$$

$$\lambda_{A'} = \frac{SP^D(v_i ,v_A^c ,G_P)}{SP^D(v_i ,v_A^c ,G_P)+SP^D(v_i ,v_{A'}^c ,G_P)}$$

such that $\lambda_A + \lambda_{A'} = 1$ and the weight for the closest insensitive node is higher.

Similarly, let the closest insensitive node to v_j in G_P be v_B^c, and the second closest insensitive node to v_j in G_P be $v_{B'}^c$, we set the weights λ_B and $\lambda_{B'}$ as

$$\text{as } \lambda_B = \frac{SP^D\left(v_j ,v_{B'}^c,G_P\right)}{SP^D(v_j ,v_B^c,G_P)+SP^D\left(v_j ,v_{B'}^c,G_P\right)} \text{ and}$$

$$\lambda_{B'} = \frac{SP^D(v_j ,v_B^c,G_B)}{SP^D(v_j ,v_B^c,G_P)+SP^D(v_j ,v_{B'}^c,G_P)}$$

such that $\lambda_B + \lambda_{B'} = 1$.

In G_Q, v_A^c, $v_{A'}^c$, v_B^c, and $v_{B'}^c$ are the centers of generalized subgraphs G_A, $G_{A'}$, G_B, and $G_{B'}$, respectively. We estimate the distance between v_i and v_j, $d(v_i,v_j)$, by integrating the estimated distances of the four possible paths going through these insensitive nodes by a linear combination with weights equal to $\lambda_a \times \lambda_b$.

$$d(v_i, v_j) = \sum_{\substack{a\in\{A,A'\} \\ b\in\{B,B'\}}} \lambda_a \times \lambda_b \times D(v_i, v_j)$$

$D(v_i, v_j)$ is the estimated distance between v_i and v_j for the path going through v_a^c and v_b^c, where a can be A or A' and b can be B or B'.

$$D(v_i, v_j) = \begin{cases} D'(G_a, v_i) + 1 + \sum_{\forall G_k}(E(G_k) + 1) + D'(G_b, v_j) & a \neq b \\ D''(v_i, v_j) & a = b \end{cases}$$

where G_k is a generalized node on the shortest path between v_i and v_j and going through v_a^c and v_b^c in a generalized graph and

If $a \neq b$ which means v_i *and* v_j are not in the same subgraph, then $D(v_i, v_j)$ is estimated by $D^{'}(G_a, v_i)$, $D^{'}(G_b, v_j)$, and $E(G_k)$. Otherwise, if v_i *and* v_j are in the same subgraph then $a = b$. In this case, $D(v_i, v_j)$ is estimated by $D^{''}(v_i, v_j)$. $D'(G_a, v_i)$ and $D'(G_b, v_j)$ correspond to the portion of the estimated distance between v_j and the subgraph gatekeeper within G_a and G_b respectively while $E(G_k)$ is the portion of the estimated distance in each subgraph G_k that the shortest path between v_i and v_j is going through in the generalized graph G_Q'.

$D^{'}(G_a, v_i)$ is computed by $E(G_a)$ and the percentage of nodes in G_a that is adjacent to the subgraph that is immediately following G_a in the shortest path between v_i and v_j in the generalized subgraph G_Q' if v_i is not the same as v_a^c. If v_i is the same as v_a^c, $D^{'}(G_a, v_i)$ is computed by the probabilities, $Prob(SP^D(v_a^c, G_a)$. $E(G_k)$ is computed by the probabilities, $Prob(SP^D(G_k)$. Computation of $D^{'}(G_b, v_i)$ is done similarly.

$$D^{'}(G_a, v_i) = \begin{cases} \left(1 - \frac{Num\ (G_a, G_k)}{Num\ (G_a)}\right) \times E(G_a) & v_i \neq v_a^c \\ \sum_{\beta = S\_SP(v_a^c, G_a)}^{L\_SP(v_a^c, G_a)} Prob(SP^D(v_a^c, G_a) = \beta) \times \beta & v_i = v_a^c \end{cases}$$

where $\frac{Num\ (G_a, G_k)}{Num\ (G_a)}$ is the percentage of nodes in G_a as a gatekeeper which is adjacent to G_k and G_k is the subgraph that is subgraph immediately following G_a in the shortest path between v_i and v_j in the generalized graph G_Q'.

$$E(G_k) = \sum_{\alpha = S\_SP(G_k)}^{L\_SP(G_k)} (Prob(SP^D(G_k) = \alpha) \times \alpha$$

$D''(v_i, v_j)$ corresponds to the estimated distance between v_i and v_j when both v_i and v_j are nodes of the same subgraph.

$$D^{''}(v_i, v_j) = \begin{cases} \sum_{\beta = S\_SP(v_a^c, G_a)}^{L\_SP(v_a^c, G_a)} Prob(SP^D(v_a^c, G_a) = \beta) \times \beta & v_i\ or\ v_j = V_a^c \\ E(G_a) & else \end{cases}$$

By using the estimated distance between any two nodes v_i and v_j, $d(v_i, v_j)$ with the shared information from G_Q', we compute the closeness centrality as follow:

$$closeness\ centrality(v_i) = \frac{n-1}{\sum_{j=1, i \neq j}^{n} d(v_i, v_j)}$$

where n is the total number of nodes in G_P

3. Experiment

We have conducted an experiment to evaluate the effectiveness of our proposed techniques. The objective is conducting aggregate social network analysis based on the incomplete information of one organization (G_p) and the shared information from another organization (G_Q).

3.1.1 Dataset - Global Salafi Jihad Terrorist Social Network

In this experiment, we used the Global Salafi Jihad terrorist social network [12],[23] as our data source to create incomplete social networks G_P and G_Q. The Global Salafi Jihad terrorist social network has 366 nodes and 1,275 links. There are four major clusters, Central Staff of al Qaeda (CSQ), Core Arab (CA), Southeast Asia (SA), and Maghreb Arab (MA). These clusters are connected in the Global Salafi Jihad terrorist social network. To simulate the real-world situation that every agency is more informative in one or two terrorist clusters but less familiar with other terrorist clusters, we randomly removed nodes from each cluster with different percentage and then randomly removed edges from the remaining subgraph. First, we generated G_P by randomly removing 30%, 30%, 70%, and 70% of nodes from CSQ, CA, SA, and MA respectively. Similarly, we generated G_Q by randomly removing 70%, 70%, 30%, and 30% of nodes from CSQ, CA, SA, and MA respectively. An edge with both of its end nodes removed was also be removed. We further removed K% of edges from G_P and G_Q. Ten pairs of G_P and G_Q were generated for each K.

3.1.2 Evaluation

In this experiment, we tested the impact of K on the performance of the proposed techniques. We also test the impact of different subgraph generalization methods, KNN and EBB, on the performance of information integration. We compared the average closeness centrality and average error obtained from the proposed techniques (E), with those obtained from the complete Global Salafi Jihad terrorist social network (G), and those obtained only from the G_P (the worst case).

3.1.3 *Experimental Result*

Table 5 and Figure 6 present the average closeness centrality of E, G, and G_P for different K by different subgraph generalization methods. Table 6 and Figure 7 present the average error in closeness centrality of E and G_P for different K by different subgraph generalization methods.

Table 5 Average Closeness Centrality of *E*, *G* and *G*$_P$ with different *K*

K	50		40		30	
	KNN	EBB	KNN	EBB	KNN	EBB
E	0.168	0.182	0.183	0.192	0.191	0.202
G	0.241		0.241		0.241	
G$_P$	0.149		0.166		0.175	

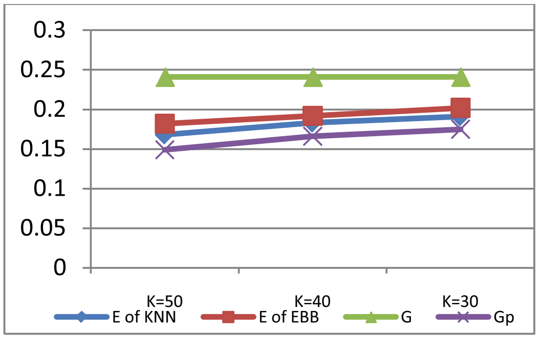

Figure 6 Average Closeness Centrality of *E*, *G* and *G*$_P$ with different *K*

Table 6 Average Error in Closeness centrality of *E* and *G*$_P$ with different *K*

K	50		40		30	
	KNN	EBB	KNN	EBB	KNN	EBB
E	0.302	0.241	0.240	0.203	0.206	0.161
G$_P$	0.378		0.312		0.274	

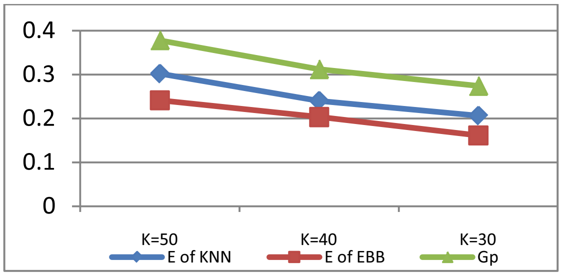

Figure 7 Average Error in Closeness Centrality of *E* and G_P with different Similarity

It is found that the error for *E* of KNN decreases from 0.302 to 0.206 when *K* decreases from 50 to 30. Similarly, the error of E of EBB decreases from 0.241 to 0.161 when K decreases from 50 to 30. The error for G_P decreases from 0.378 to 0.274 when *K* decreases from 50 to 30. The average error of E of KNN is consistently lower than that of G_P by a substantial amount. But the average error of E of EBB is even lower consistently than E of KNN. When K = 50, 40, and 30, the error of *E* of KNN is reduced from the error of G_P by 20%, 23%, and 25% respectively. The error of E of EBB is reduced from the error of G_P by 37%, 35%, and 41% respectively. It shows that more accurate closeness centrality can be obtained by integrating the generalized information from G_Q with G_P. Further, different subgraph generalization methods have significant impact on the final result of information integration.

4. Conclusion

Increasing number of social network data has been made publicly available and analyze. In intelligence and security informatics domain, social network analysis is very useful in investigating the terrorist and criminal communication patterns and the structure of their organizations. However, most law enforcement force and intelligence agents only have a small piece of information before integration with the information from other agents. Further, information sharing is not always possible due to privacy issue. In this paper, we propose to construct generalized graphs before sharing the social network with other parties. The generalized graph will then be integrated with the social network owned by the agent himself to conduct social network analysis such as closeness centrality. By sharing and integrating generalized information, an agent not only preserves the privacy of data but also preserves the utility of social network data. Our experiment shows that our proposed techniques improve the closeness centrality measurements substantially. Further, the experimental results show that different subgraph generalization methods can make significant impact on the effectiveness of information integration. In our work, using edge betweenness based method to generalize social network yields consistently lower error rate in closeness computation than using k-nearest neighbor method. In our future work, we shall extend our work to integrate multiple instead of two networks. We will also try to formulate this problem into an optimization process with some graph spectral features.

5. REFERENCES

[1] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography," in *WWW'07* Banff, Alberta, Canada, 2007.

[2] Z. Baird, J. Barksdale, and M. Vatis, Creating a Trusted Network for Homeland Security, Markle Foundation, 2003.

[3] Z. Baird and J. Barksdale, Mobilizing Information to Prevent Terrorism: Accelerating Development of a Trusted Information Sharing Environment, Markle Foundation, 2006.

[4] K. Caruson, S. A. Macmanus, M. Khoen, and T. A. Watson, "Homeland Security Preparedness: The Rebirth of Regionalism," *Publius*, 35(1), 2005, pp.143-189.

[5] R. R. Friedmann and W. J. Cannon, "Homeland Security and Community Policing: Competing or Complementing Public Safety Policies," *Journal of Homeland Security and Emergency Management*, 4(4), 2005, pp.1-20.

[6] M. Girvan and M. Newman, Community structure in social and biological networks, Proceedings of the National Academy of Sciences, 99 (2002), pp. 7821..

[7] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava, "Anonymizing Social Networks," Technical Report 07-19, University of Massachusetts, Amherst 2007.

[8] N. Li and T. Li, "*t*-closeness: Privacy Beyond k-anonymity and l-diversity," in *ICDE'07*, 2007.

[9] K. Liu and E. Terzi, "Towards Identity Anonymization on Graphs," in *ACM SIGMOD'08* Vancouver, BC, Canada: ACM Press, 2008.

[10] Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-Diversity: Privacy Beyond k-Anonmity," Proceedings of the 22nd International Conference on Data Engineering, 2006.

[11] M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the Presence of Individuals from Shared Database," in *SIGMOD'07*, 2007.

[12] M. Sageman, *Understanding Terror Networks*, University of Pennsylvania Press, 2004.

[13] P. Samarati, "Protecting Respondents' Identities in Microdata Release," IEEE Transactions on Knowledge and Data Engineering, 2001.

[14] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," International Journal on Uncertainty Fuzziness Knowledge-based Systems, 10(5), 2002, pp.557-570.

[15] D. Thacher, "The Local Role in Homeland Security," *Law & Society*, 39(3), 2005, pp.557-570.

[16] B. Thuraisingham, "Security Issues for Federated Databases Systems," Computers and Security, North Holland, December, 1994.

[17] B. Thuraisingham, Chapter 1. Assured Information Sharing: Technologies: Challenges and Directions, Intelligence and Security Informatics: Applications and Technique, Editors: H. Chen and C. C. Yang, Springer-Verlag, to appear in 2008.

[18] R. C. Wong, J. Li, A. Fu, and K. Wang, "(α,k)-Anonymity: An Enhanced k-Anonymity Model for Privacy-Preserving Data Publishing," Proceedings of SIGKDD, August 20-23, Philadelphia, Pennsylvania, US, 2006.

[19] X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proceedings of SIGMOD, June 27-29, Chicago, Illinois, 2006.

[20] X. Xiao and Y. Tao, "m-invariance: Towards Privacy Preserving Re-publication of Dynamic Datasets," in ACM SIGMOD'07: ACM Press, 2007.

[21] C. C. Yang, N. Liu, and M. Sageman, "Analyzing the Terrorist Social Networks with Visualization Tools," Proceedings of the IEEE International Conference on Intelligence and Security Informatics, San Diego, CA, US, May 23 – 24, 2006.

[22] C. C. Yang, "Information Sharing and Privacy Protection of Terrorist or Criminal Social Networks," *Proceeding of IEEE International Conference on Intelligence and Security Informatics*, Taipei, Taiwan, 2008.

[23] C. C. Yang and M. Sageman, "Analysis of Terrorist Social Networks with Fractal Views," *Journal of Information Science*, accepted for publication.

[24] X. Ying and X. Wu, "Randomizing Social Networks: A Spectrum Preserving Approach," in *SIAM International Conference on Data Mining (SDM'08)* Atlanta, GA, 2008

[25] B. Zhou and J. Pei, "Preserving Privacy in Social Networks against Neighborhood Attacks," in *IEEE International Conference on Data Engineering*, 2008