Vol.2015-CSEC-69 No.31
Vol.2015-IOT-29 No.31
2015/5/22

SNSの不正アカウント検出においてコミュニティ構造に着目したグラフ剪定及びシード選択法

春田　秀一郎[†1,a)]　豊田　健太郎[†1,b)]　笹瀬　巌[†1,c)]

概要：近年, SNS (Social Networking Service) において, スパムを配信する等の行為を行う不正アカウントの検出が急務である. その検出法として, 友人数の多いシードと呼ばれる代表アカウントを起点にその友人に信頼値を分配し, 信頼値の低い者を不正アカウントとして検知する PI (Power Iteration) 法及び, 不正アカウントに流入する信頼値を抑制するため, 共通の友人が少ない者同士の友人関係を不正アカウント-正規アカウント間の友人関係と見なして関係を剪定する GP (Graph Pruning) 法が存在する. しかし, PI 法においては, 一般に友人数の多いアカウントは同一コミュニティに属す傾向があるため, 選択されるシードが偏り, 信頼値が均一に分配されない問題がある. また攻撃者が複数の不正アカウントを用いることで共通の友人数を増大し, GP 法における剪定を回避できる問題がある. そこで本稿では正規アカウントに対してより均一に信頼値を割り当てるため, SNS 全体に対してコミュニティ検出を行い, 検出された各コミュニティの中から友人数の多いアカウントをシードとして選択する方式を提案する. さらに選択したシードを起点に信頼できるアカウントの領域を求めることで複数の不正アカウントを用いた場合に対してもロバストな剪定方式を提案する. これら 2 つの提案により, 正規アカウントに分配される信頼値を増大し, 不正アカウントに分配される信頼値を低減することを可能とする. 実データを用いた特性評価を行い, 提案方式は従来方式と比較して各アカウントの正規性をより正確に判別可能であることを示す.

1. はじめに

近年, スマートフォンが身近なものとなると共に, Facebook や Twitter をはじめとした SNS (Social Networking Services) が普及し, SNS を利用することでユーザはインターネット上で交流することが可能となった. このような便利な側面もある一方, 悪意のあるユーザが多量の不正アカウントを生成し, そのアカウントから正規アカウントにスパムを送信したり, 投票型システムに対して不正票を大量に投じる等といった Sybil 攻撃が問題視されている. したがって, サービス利用者が安心して SNS を利用できるために, システムは不正アカウントを検出する必要がある.

近年, 不正アカウント検出法の 1 つに PI (Power Iteration) 法が注目されている [1]. この手法では友人数の多いシードと呼ばれる代表アカウントに信頼値を定数として与え, このシードを起点としてその友人に信頼値を等分配し, さらにシードの各友人を起点として分配された信頼値をその友人に対して分配するという操作を一定数繰り返す. そしてこの操作を複数のシードに対して行い, 最終的に信頼値の低い者を不正アカウントとして検知する. ここで, 不正アカウント-正規アカウント間の友人関係は AE (Attack Edge) と呼ばれ, この AE を通して不正アカウントへ信頼値が分配されることは問題なので, これを防止するために, PI 法に先立って共通の友人が少ない者同士の友人関係を AE と見なして剪定する GP (Graph Pruning) 法 [2] が存在する.

しかしながら, 単純に PI 法を適用した場合, 正規アカウントを不正アカウントとして判別する偽陽性が発生する. また, GP 法においては悪意のあるユーザは複数の不正アカウントを用いることで共通の友人数を増大させることが可能であり, 共通の友人数は不正アカウントを複数用いることで増加することが可能であるため, 共通の友人数を剪定基準に用いることは適切でないと考えられる.

そこで本稿では, まず 1 つ目の問題点は, 全正規アカウントに適切に信頼値が分配されないことに原因があると考え, 実際の SNS のデータセットを用いて, 友人数の多いアカウントが SNS 内においてどのように分布するかについて調

[†1] 現在, 慶應義塾大学 理工学部情報工学科
Presently with Dept of Information and Computer Science Keio University
[a)] haruta@sasase.ics.keio.ac.jp
[b)] toyoda@sasase.ics.keio.ac.jp
[c)] sasase@ics.keio.ac.jp

ⓒ 2015 Information Processing Society of Japan

べた. その結果, 一般に友人数の多いアカウントが同一のコミュニティに偏る傾向があり, 同一のコミュニティからシードを選択するため, 他のコミュニティに属すアカウントに適切に信頼値が分配されないことが分かった. そこで提案方式では, SNS 全体に対し, 既存のコミュニティ検出法を用い, 検出した各コミュニティの中で友人数の多いアカウントをシードとして選択することで, 正規アカウントに対してより均一に信頼値を割り当てることを可能とする. さらに, GP 法において共通友人数を剪定基準とせず, 選択したシードを起点に信頼できるアカウントの領域を求め, その領域の境界に存在する友人関係を剪定する方式を提案する. これにより悪意のあるユーザが複数の不正アカウントを用いた場合においてもロバストな剪定を可能とする. 以下 2 章では本研究の背景を, 3 章では関連研究と従来方式について, 4 章では提案方式について説明する. さらに 5 章では提案方式の有効性を示すため, 計算機シミュレーションを用いて判別アルゴリズムの性能についての特性評価を行う. 最後に 6 章で本稿のまとめと今後の課題について述べる.

2. 背景

2.1 システムモデル

本研究における SNS は正規アカウント数 n_H と不正アカウント数 n_S からなる全アカウント数 $n = n_H + n_S$, 全友人関係数 m の無向 SNS を想定する. ここで全友人関係数とは, あるアカウント間が友人関係にある場合を 1 とカウントした際にネットワーク全体で m の友人関係があることを表す. 無向 SNS とは, Facebook のように, 相互の承認があって初めて友人関係が構築されるような SNS を指す. SNS のサービス事業者は, 全 n アカウントおよび各アカウント間の友人関係を把握しているが, どのアカウントが不正であるかは不明であり, サービス事業者は SNS 上の正規アカウントを不正アカウントと誤判別することなく n_S 個の不正アカウントを可能な限り多く検知することを目的とする.

2.2 攻撃モデル

本研究では複数の攻撃者 $\mathcal{A} = \{a_i\}$ ただし, $1 \leq i \leq n_{att}$ が存在し, 正規アカウントにスパムを配信する等の Sybil 攻撃を行うことを想定する. 各攻撃者はシステムに対して以下のような操作が可能であると仮定する.

(1) n_S 個の不正アカウントを生成
(2) 生成した任意の不正アカウント間において友人関係を形成
(3) 正規アカウントとの間に m_{AE} 個の友人関係 (AE:Attack edge) を形成

上記 (2) の操作により，不正アカウント間が正規アカウントと同等の友人関係を形成することを可能とする．(3) に関して, m_{AE} は SNS 全体における友人関係数 m に比べて小さな値となる．これは, 正規アカウントは未知のアカウントからの友人申請を許可しにくいためである．図 1 に本研究における SNS の概略図を示す.

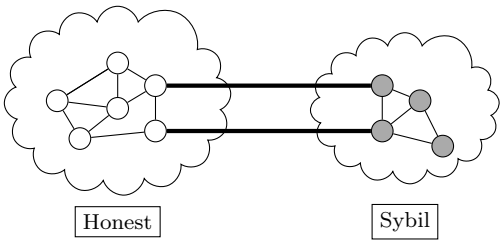

図 1 正規アカウントコミュニティと不正アカウントコミュニティを結ぶ AE の関係. 太線に示すのが AE である

3. 従来方式

3.1 関連研究

近年の Sybil 攻撃に対する手法として, [3], [4] がある. 不正アカウントは大量に作られる一方, AE の数は少数であることを仮定し, ネットワーク上で有限回数のランダムウォークを行う. 各アカウントがランダムウォークを行い, verifier と呼ばれるアカウントからのランダムウォークのルートが一致したものを正規アカウントとする. この考えに基づいて不正アカウントの検出を行うが, これらの手法は偽陽性及び偽陰性を多く含む. さらに, ランダムウォークのランダム性を低減した手法として, 本研究では, 手法に着目する.

3.2 Power Iteration 法

Sybil 攻撃によって生成された不正アカウントの検出手法として PI 法 [1] がある. この手法では友人数の多いシードと呼ばれる代表アカウントに信頼値を定数として与え, このシードを起点としてその友人に信頼値を分配し, さらにシードの各友人を起点として分配された信頼値をその友人に対して分配するという操作を 1 ラウンドとして複数ラウンド繰り返す. 各アカウントは分配された信頼値を加算していく. 最終的に各アカウントの信頼値をあらかじめ定めた閾値と比較することで, 各アカウントを正規アカウントもしくは不正アカウントに分類する．以下により詳細なアルゴリズムを示す．シードの数を M としたとき, 各シードに信頼値 T_G を等分配し, その他のアカウントの信頼値は 0 とする. すなわち, アカウントの信頼値の初期値 $T^{(0)}(v)$ は式 (1) として表される.

$$T^{(0)}(v) = \begin{cases} \frac{T_G}{M} & \text{If } v \text{ is a seed.} \\ 0 & \text{Otherwise} \end{cases} \quad (1)$$

各シード v は v の友人アカウントに信頼値 $T^{(0)}(v)$ を等分配する．すなわち，友人アカウント u の信頼値 $T^{(1)}(u)$ は,

$$T^{(1)}(u) = \sum_{u_j \in U_v} \frac{T^{(0)}(v)}{deg(u_j)} \qquad (2)$$

となる. ただし, U_v を v の友人の集合, $deg(v)$ を v の友人数とする. 以上の操作を各シードの友人に対して再帰的に $i = \lceil \log n \rceil$ 回繰り返す. そして最終的な各アカウントの信頼値 $T^{(i)}(u)$ を降順に並べ, 下位のアカウントを不正アカウントとして判別する [1]. PI 法の利点は $O(n \log n)$ の計算量でアカウントの正規性の判別が可能である点にある. この $O(n \log n)$ は信頼値の分配に $O(n)$, 信頼値の分配に $O(\log n)$ の計算量がかかることから算出される.

3.3 Graph Pruning 法

GP (Graph Pruning) 法 [2] は, PI 法を適用する前に予め AE の可能性が高い友人関係を剪定しておくことで不正アカウントへ分配される信頼値を低減し, 不正アカウント及び正規アカウントの判別の正確性を向上する. 共通の友人が少ない者同士の友人関係は AE である可能性が高いという仮定に基づき, アカウントとアカウントの共通友人数が閾値 T_s より小さい場合にその友人関係を剪定する.[2] において, 閾値は $T_s = 1$ を用いている.

3.4 従来方式の問題点

3.4.1 PI 法の問題点

PI 法では, シードを友人数の多いアカウントの中からランダムに選択する. しかし, 一般に友人数の多いアカウント同士は同一コミュニティに属すため, 選択されるシードが偏る傾向がある. 図 2 は, [5] にて公開されているアカウント数 4039 の Facebook のデータセットにおいて, 全アカウントを友人数の降順で並べた場合に上位 K% のアカウントがどのコミュニティに属しているかを表す. コミュニティ検出法には FG (Fast Greedy 法)[6] を用いた. 例えば, $K = 5$% のとき, 友人数上位 5% のアカウントは 3~4 個のコミュニティのいずれかに属していることがわかる. 図 2 より, K が小さい時, 友人数の多いアカウントは少数のコミュニティに集中していることが分かる. $K = 30$% においても, 全 13 コミュニティのうち 5 のコミュニティに集中していることが分かる. 図 3 に PI 法で選択されるシードの例を示す. このように単純に友人数の多いアカウントからシードを選択した場合, 同一コミュニティに属するアカウントをシードとして選択する可能性が高くなる. その結果, シードから離れた場所に位置する正規アカウントに十分な信頼値が分配されず, 不正アカウントと判断されてしまう問題点がある.

3.4.2 GP 法の問題点

GP 法では共通友人数をもとに剪定する友人関係を選択するが, 攻撃者は容易に共通友人数を増加させることが可能である. 図 4 に, 攻撃者がある正規アカウントに対し, 複数の AE を構築した後に不正アカウント間で友人関係を構築することで共通友人数を増大させる例を示す. 白色のアカウント H は正規アカウント, 灰色のアカウント S_1, S_2, S_3, S_4 は SN であり, 太線は AE を示す. SNS において未知のアカウントからの友人申請を許可するユーザはどのようなアカウントからの友人申請も承認する可能性が高く, 複数の不正アカウント S_i からの友人申請を許諾し得る. そこでまず攻撃者は AE を受容するアカウントを探索し, 発見次第, 複数の不正アカウントを用いて対象の正規アカウントに対し複数の AE を構築する. その後, S_1-S_4 は各々 H と AE を構築した後, S_i 間で友人関係を結ぶ. 攻撃者は S_1-S_4 を管理しているため, この操作を容易に行うことができる. その結果, 対象の正規アカウントと不正アカウント間の共通友人数は 3 となり, 剪定基準である $T_s < 1$ から逸脱することが可能である. 従って共通友人数を基準とした GP 法では攻撃者が図 4 のような AE を切断できないため, 不正アカウント及び正規アカウントの判別の正確性

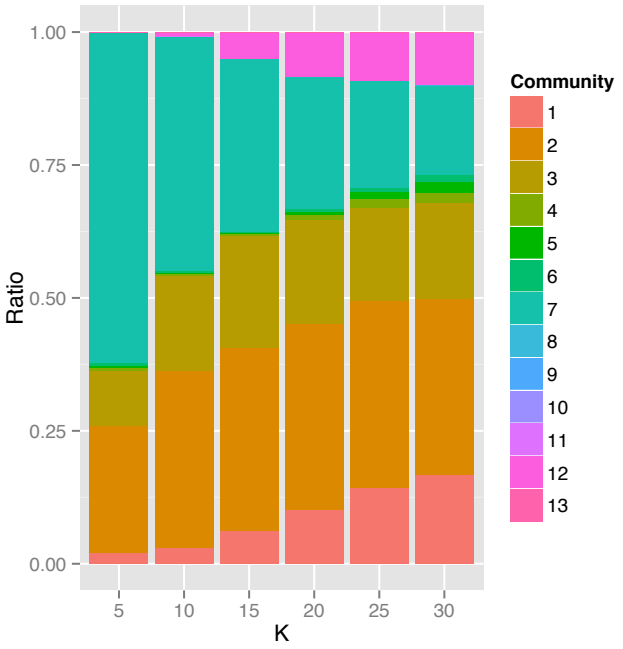

図 2 各コミュニティに含まれる友人の多いアカウントの割合

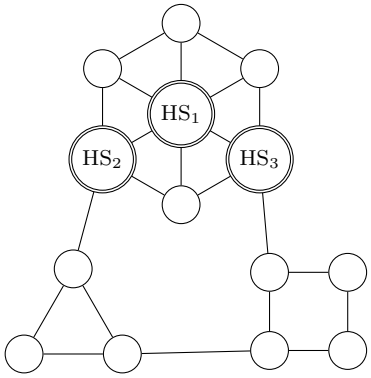

図 3 従来方式で選択されるシードの例

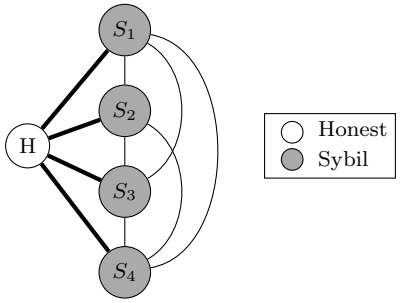

図 4　SN 側から共通友人数を増加させる例

が低減するという問題点がある.

4. 提案方式

提案方式では, 上記 2 つの問題点を解決するために, コミュニティ構造を考慮したシード選択方法および信頼領域を基にした AE の剪定方法を提案する. 1 つ目の提案手法は, PI 法におけるシード選択時に, コミュニティ検知アルゴリズムを用いて SNS のコミュニティ構造を把握しておき, 各コミュニティ内において友人数の多いアカウントをシードとして選択する. これにより 1 つ目の問題点であった, 正規アカウントへの信頼値の分配が属するコミュニティによって偏る問題点を解決する.2 つ目の提案手法は, 選択されたシードおよびその友人を初期の信頼領域とし, その周辺の友人が信頼領域内の友人と一定以上友人関係がある場合のみその友人を信頼領域に取り込み, そうでない場合は確率を用いて剪定を行う. この剪定確率は信頼領域内のアカウントに対して構築された友人関係数が少ない程低くする. これにより, 例え攻撃者が不正アカウント間で友人関係を構築したとしても AE を剪定することを可能とする. 提案方式では上記 2 つの問題点を改善することにより, アカウントの正規性の判別精度を向上する.

4.1 アルゴリズム

まず, FG 法 [6] を用いて, コミュニティ検出を行い, 各コミュニティから友人数が全体の SNS に対して上位 K% のアカウントをシードとして選択する. あるコミュニティにこのようなアカウントが存在しない場合, そのコミュニティからシードは選択しない. SNS のような大規模なネットワークに対するコミュニティ検出には計算コストの少ないコミュニティ検出法が有効であり, FG 法は $O(n\log^2 n)$ という少ない計算量でコミュニティ検出が可能であるためこの手法を選択した. 図 5 提案方式におけるシードの選択例を示す. 図 5 において FG 法により, 3 つのコミュニティ A, B, および C が検出されたとし, 各コミュニティから 1 つずつシードを選択する. 例えばアカウント A_1 から A_7 からなるコミュニティ A において, 最も友人数の多いアカウントは A_4 である. したがって A_4 をコミュニティ A のシードとして選択する. 同様の操作をコミュニティ B

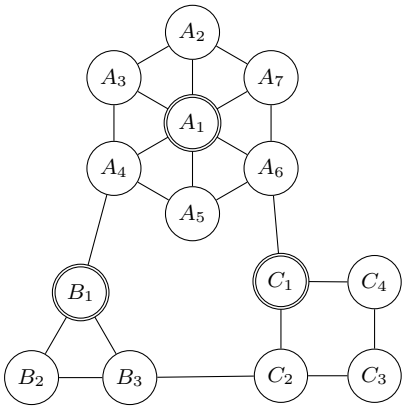

図 5　提案方式で選択されるシードの例

および C に対しても行う. このとき最も友人数が多いシードが複数存在する場合はランダムに 1 つ選択する. 例えば B_1, B_2, および B_3 からなるコミュニティ B において最も友人数が多いアカウントは B_1 と B_3 であるが (共に友人数 3), この例ではランダムに B_1 が選ばれたものとする.

4.2 信頼領域に基づく AE 剪定法

前節で求めた各シードを元に, AE を剪定する. まず各シードおよびそれらの直接の友人を初期の信頼領域 TA (Trusted Area) とする. 次に TA に隣接する各アカウントを TA に追加するかを決定する. ここで, TA への追加基準として, 評価対象のアカウントがどの程度 TA 内のアカウントと友人関係にあるかを用いる. そこで評価対象アカウント u の友人数を $deg(u)$, そのうち TA 内に存在するアカウントとの友人関係数を $deg(u)_{\text{in}}$ とした時, u の TA 追加指数 $T_{\text{TA}}(u)$ を

$$T_{\text{TA}}(u) = \frac{deg(u)_{\text{in}}}{deg(u)} \tag{3}$$

と定める. ここで R_{TH} を TA 追加指数の閾値とし,$T_{\text{TA}}(u) \geq R_{\text{TH}}$ の場合, このアカウントを TA に追加する. R_{TH} はヒューリスティックな値として $R_{\text{TH}} = \frac{2}{3}$ を用いる. 上記の操作を TA に含むことが可能なアカウントが存在しなくなるまで行う. TA 内のアカウントと TA に含まれなかったアカウントとの間の友人関係は AE の可能性がある. このとき, $T_{\text{TA}}(u)$ の値が低い程, TA 内のアカウントとの関係が希薄であるため, AE の可能性が高いと言える. そこで, TA に含まれなかったアカウントと TA に追加されたアカウントの間にある友人関係を, TA 追加指数に応じた確率を用いて友人関係を剪定する. TA 追加指数を用いて, その友人関係を切断する確率 P_{cut} を,

$$P_{cut} = 1 - \frac{T_{\text{TA}}(u)}{R_{\text{TH}}} \tag{4}$$

とする.

図 6 に上記剪定法の例を示す. 白色のアカウントは既に TA に取り込まれたアカウント, 灰色のアカウントは現在

ⓒ 2015 Information Processing Society of Japan

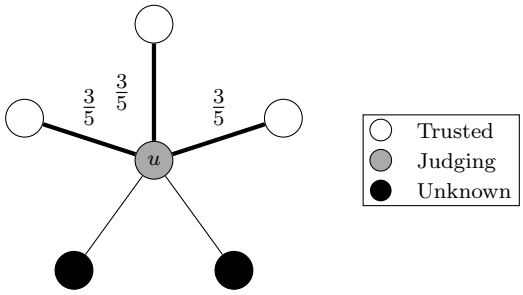

図 **6** アカウント u の TA への取り込みを行うかの判定例

評価対象のアカウント, 黒色のアカウントは現在 TA に取り込まれていないアカウントを示す. 図 6 において白色のアカウントと灰色のアカウントで結ばれた線上の値は u の TA 追加指数 $T_{\mathrm{TA}}(u)$ を表している. このときアカウント u の友人数は $deg(u)(=5)$, アカウント u の友人のうち, 信頼されているアカウントは $deg_{in}(u)(=3)$ であり, 評価対象アカウント u は $T_{\mathrm{TA}}(u) = \frac{3}{5} < \frac{2}{3} = R_{\mathrm{TH}}$ となるため, TA 追加の基準である $T_{\mathrm{TA}}(u) \geq R_{\mathrm{TH}}$ を満たさない. したがって, u は TA に追加されず, u と白色のアカウントを結ぶ友人関係の剪定確率 P_{cut} は, $P_{cut} = 1 - (3/5)/(2/3) = \frac{1}{10}$ となる.u が TA に取り込まれなかったため, 黒色のアカウントには TA 追加の評価は行われない.
上記の提案方式を適用後, 本方式では PI 法を用いて不正アカウントを検出する.

5. 特性評価

提案方式の有効性を示すため, 以下の二つの攻撃モデルに対して, 判別アルゴリズムの性能を示す ROC (Receiver Operating Characteristic) 曲線の AUC (Area Under Curve) の値を評価した.AUC の値は 1 に近いほど良質な判別アルゴリズムであることを示す. 以下, 不正アカウントから AE を追加される正規アカウントを SSR (Sybil Supporter) と呼ぶ. また, データセットには Facebook[5] を用い, アカウントが形成するコミュニティは平均友人数 10 のランダムグラフ (BA モデル [8]) とした. 表 1 に, シミュレーションに用いたパラメータをまとめる.

表 **1** パラメータとその値

パラメータ	値
K	5
R_{TH}	0.66
T_s	1
T_p	2

5.1 攻撃モデル

5.1.1 攻撃モデル 1:従来の攻撃方法

攻撃モデル 1 ではランダムに選択した 100 個の SSR に対して, ランダムに選択した不正アカウントから合計 $m_{\mathrm{AE}} = 200$ の AE を追加する. 図 7 に攻撃モデル 1 の概略図を示す.

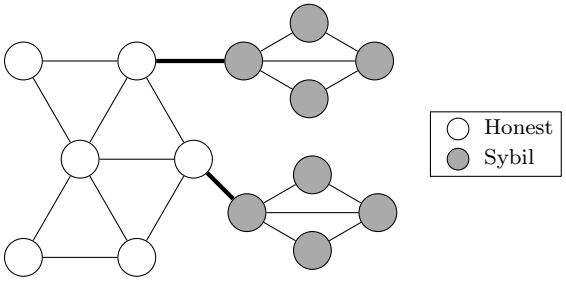

図 **7** 攻撃モデル 1 の概略

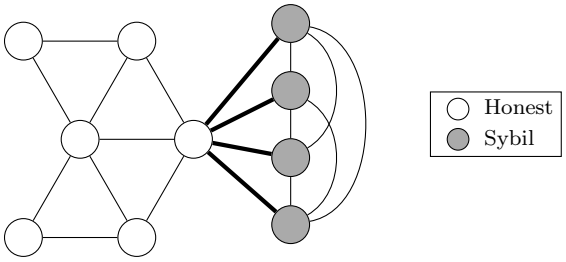

図 **8** 攻撃モデル 2 の概略

5.1.2 攻撃モデル 2:一つのユーザに集中的に AE を追加

攻撃モデル 2 ランダムに選択した各 20 個の SSR に対して, 一つの SSR 毎にランダムに選択した 10 個の不正アカウントから AE を追加し, 合計 $m_{\mathrm{AE}} = 200$ の AE を追加した上で, 共通友人数を増加させるため, 一つの SSR に AE を追加した不正アカウント同士に友人関係を持たせた. 図 8 に攻撃モデル 2 の概略図を示す.

5.2 評価

5.2.1 攻撃モデル 1

図 9 及び, 図 10 に攻撃モデル 1 においてそれぞれ $n_{att} = 1$, $n_{att} = 5$ とした際の不正アカウントの数と AUC の変化を示す. $n_{att} = 1, n_{att} = 5$ のいずれの場合においても提案方式は従来方式と同程度の AUC を実現している. 攻撃モデル 1 においては, 多くの SSR に対して AE を追加するため, GP 法による友人関係の剪定が効果的に作用する. そのため多くの AE を剪定することができたと考えられ, 高い AUC の値を保ったと考えられる.

5.2.2 攻撃モデル 2

図 11 及び, 図 12 に攻撃モデル 1 においてそれぞれ $n_{att} = 1$, $n_{att} = 5$ とした際の不正アカウントの数と AUC の変化を示す. 攻撃モデル 2 では, 従来方式は AUC の値が著しく低減した. これは従来方式では共通友人数に着目して友人関係を剪定するため, 本攻撃モデルでは友人関係を剪定できないためである. 提案方式では効果的に AE が剪定されたことで AUC を高く保つことができている.

ⓒ 2015 Information Processing Society of Japan

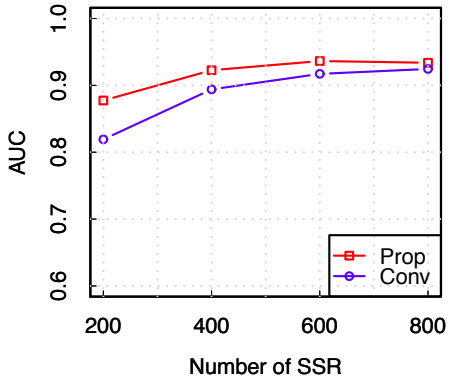

図 9　攻撃モデル 1 :$n_{att} = 1$

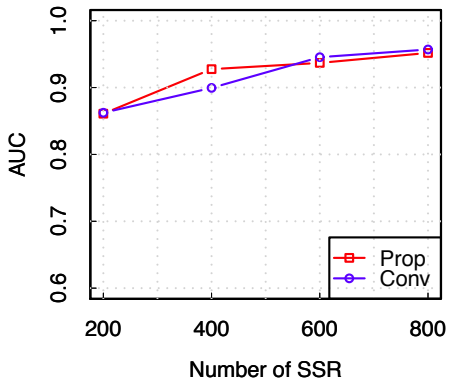

図 10　攻撃モデル 1 :$n_{att} = 5$

6. おわりに

本稿では, 近年 SNS 上で問題視されている不正アカウントの検出法として PI 法に着目し, 従来のシード選択方法では選択されるシードが一つのコミュニティに偏る傾向があることを示した. また GP 法においては, 従来の剪定基準では攻撃者が容易に剪定を回避できることを示した. 以上の 2 つの問題点に対してコミュニティ構造に着目したシード選択法と, 信頼領域を用いた友人関係の剪定法を提案し, 計算機シミュレーションにより判別アルゴリズムの性能を示す AUC について評価し, 提案方式の有効性を示した.
今後は初期にシードに与える信頼値をコミュニティの大きさに応じて変化させることや, 詳細な特性評価を行っていく予定である.

謝辞　本研究の一部は, 「科研費 基盤研究 (C)26420369 高信頼性を有するＩｏＴの実現に向けたセキュアアクセス制御方式に関する研究」の助成により行われた. 関係者各位に深謝する.

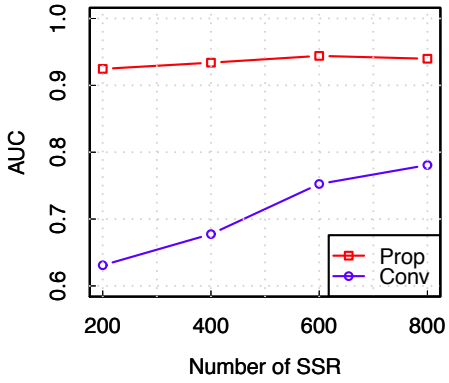

図 11　攻撃モデル 2 :$n_{att} = 1$

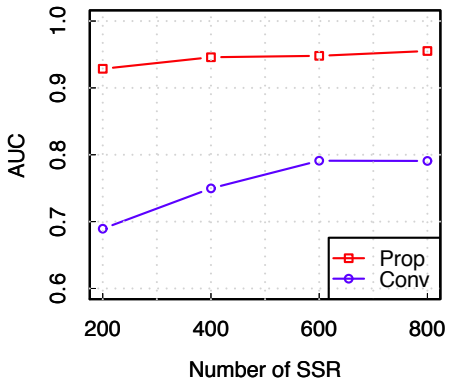

図 12　攻撃モデル 2 :$n_{att} = 5$

参考文献

[1] Cao, Qiang et.al.: Aiding the Detection of Fake Accounts in Large Scale Social Online Services, USENIX NSDI (2012).
[2] Zhang, Huanhuan et.al.: Exploiting Trust and Distrust Information to Combat Sybil Attack in Online Social, Trust Management VIII (2014).
[3] Yu, Haifeng, et al: Sybilguard: defending against sybil attacks via social networks, ACM SIGCOMM Computer Communication Review. Vol. 36.4 (2006)
[4] Yu, Haifeng, et al: Sybillimit: A near-optimal social network defense against sybil attacks, Security and Privacy IEEE Symposium (2008)
[5] Stanford Network Analysis Project：Social circles: Facebook. 入手先 ⟨https://snap.stanford.edu/data/egonets-Facebook.html⟩ (2015.03.20).
[6] Newman et.al.: Fast algorithm for detecting community structure in networks, Physical review E 69.6 (2004)
[7] Hanley et.al.: he meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology vol.143.1 p29-36 (1982)
[8] Barabashi et.al: Emergence of scaling in random networks, science vol.286.5439 p509-512 (1999)
[9] Behrends, E: Introduction to Markov Chains, with Special Emphasis on Rapid Mixing, Vieweg & Sohn, Braunschweig, Wiesbaden, Germany (2000)

ⓒ 2015 Information Processing Society of Japan