The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

多様なソーシャルネットワーク構造を用いたcyber-predator予測分析

Predictive Analysis on Cyber-predators Using Various Social Network Structures

西口　真央 *1　　鳥海　不二夫 *1
Mao Nishiguchi　　Fujio Toriumi

*1東京大学
The University of Tokyo

In this paper, we use the information of social network structures to tackle cyber-predator detection in a social networking service, and compare and analysis the explanatory power of these structures. We first create networks from various perspectives such as footprints and reactions to posts as well as conversations. By applying Large-scale Information Network Embedding (LINE) to these networks, latent representations based on each network structure are extracted. Using these latent representations as input features, we develop classification models for predicting cyber-predators. As a result of computational experiments, we confirmed that many social network structures are effective for detecting cyber-predators. In addition, we got some interesting findings, such as "the tendency of cyber-predator appears most strongly in the profile browsing history". The findings obtained in this paper are used to suppress minors' cybercrime damage.

1. はじめに

近年，オンラインコミュニティを介した未成年者の誘い出しやいじめなどの犯罪被害が増加傾向にある．警察庁の調査によると，特に複数人での交流が可能なソーシャルネットワーキングサービス（SNS）に起因する被害児童数が顕著に増加しており，その数は過去最多となっている [警察 17]．スマートフォンや SNS を日常的に利用する事が一般的となった現代において，サイバー犯罪の加害者，いわゆる cyber-predator を正確に検知し上記のような事犯を限りなく抑えることが，喫緊の社会的課題の 1 つとなっている．

これまでにも，cyber-predator をシステム的に検知するための研究は数多く行われてきた [Nahar 13, Huang 14, F. Toriumi 15, Cardei 17, Liu 17]．cyber-predator 検知には，会話から抽出されたテキストコーパスを入力とするものが多いが，コーパスのみを用いたモデルでは予測性能に限界がみられる [Huang 18]．近年では，テキストコーパス以外にも，ユーザ属性やユーザ間の会話の有無から得られるネットワーク構造を活用する手法も盛んに提案されており，これらの入力は検知精度の向上に貢献することが実験的に示されている [Nahar 13, Huang 14, F. Toriumi 15, Cardei 17].

しかしながら，既存研究の多くは事後的，すなわち cyber-predator との接触後に検知することを想定したものであり，接触自体を防ぐことは難しい．サイバー犯罪を抑えるためには，cyber-predator をいち早く検知し，接触を未然に防ぐ仕組み作りもまた重要である．また，近年の主要な SNS におけるユーザ同士の繋がりは，会話の他にもフォローや通報，ブロックなど，様々な定義のネットワークが考えられるが，これらのネットワーク構造を活用した cyber-predator 検知に関する研究は，我々が知る限りまだ無い．

以上の背景を踏まえ，本研究では，cyber-predator 未然検知システムを構築する第 1 歩として，未成年者が多く存在する複数交流形 SNS 内に蓄積された実データを利用し，様々な種類のソーシャルネットワーク構造の有用性を比較分析する．ソーシャルネットワークは一般的に大規模かつ疎であり，ネットワーク構造をそのまま活用することは困難である．そこで我々は，[Tang 15] によって提案された Large-scale Information Network Embedding (LINE) により，ネットワーク構造を低次元のベクトル空間で表現することを試みる．LINE は，エッジの重みや向きも考慮可能な，大規模ネットワーク構造に適した分散表現獲得手法の 1 つである．我々はまず，各種ネットワーク構造から分散表現を独立に学習する．次に，cyber-predator を予測するための適切なクラス分類問題を設定し，それぞれの特徴量を入力としたモデルを個別に構築することで，各種ネットワーク構造の有用性を比較検証する．

次章では，cyber-predator に関連する既存研究について紹介する．続く第 3 章では，分散表現獲得手法である LINE を説明する．第 4 章では計算実験の設定及び結果について議論する．最後に，本研究で得られた知見と今後の課題について述べる．

連絡先: 西口　真央，東京大学，東京都文京区本郷 7-3-1，03-5841-6991，nishiguchi@crimson.q.t.u-tokyo.ac.jp

2. 関連研究

本章では，大規模データからの cyber-predator の自動検知に関する既存研究について議論する．検知モデルの入力に最も利用されてきたデータセットは，ユーザ同士の会話内容から抽出されたテキストコーパスである．テキストコーパスは，モデルに対する強い説明力である一方，Bag of Words などの単純な特徴だけでは検知精度に限界がある [Huang 18]．そのため，近年ではコーパス以外のデータも利用する研究も盛んに行われている．

SNS に蓄積されたコーパス以外のデータセットの中で，注目されていものの 1 つが，ユーザ間の繋がりを表すソーシャルネットワークデータである．[Nahar 13] は，過去にサイバー空間上で行われたいじめの被害者と加害者を一種のソーシャルネットワークとして表現し，次に最も活発なサイバーいじめが行われるリンクを識別する手法を開発した．ユーザ間のソーシャルネットワークが，サイバー犯罪の検知に寄与することは，その他の研究でも示唆されている．[Huang 14] は，テキストコーパスに加え，[Nahar 13] と同様にサーバー空間におけるいじめの加害者-被害者ネットワークを構築し，独自に定義したネットワーク構造の分散表現などをモデルの入力として

与えることで，予測性能の改善に成功した．[F. Toriumi 15] では，ユーザ間の会話をエッジとみなした有向ネットワークを作成し，ネットワーク構造に基づくクラスタリング分析を行ったところ，誘い出しを受けるユーザの構造と誘い出す側のユーザとの間に，明確な違いがあることを確認した．

これらの既存研究は，ソーシャルネットワークが cyber-predator 検知に効果的であることを示した価値のある研究であるが，我々は大きく 2 つの課題が残されていると考える．1 つ目は，すでに事犯が発生した後のネットワーク構造を利用している点である．犯罪被害をさらに抑えるためには，cyber-predator である可能性が高い者との接触自体を未然に防ぐことが望ましい．2 つ目は，既存の研究で定義されたソーシャルネットワークが，直接的な会話の有無により定義されるものに限定されていることである．現在の主要な SNS には，ユーザ同士の会話以外に，フォローやコメント，リツイートといった機能があり，これらも 1 種のソーシャル関係を意味する．ネガティブなものでは通報やブロックといった繋がりも，ソーシャルネットワークとして捉えることができる．

したがって，本研究では，事犯が発生するより以前のデータのみを利用し，検知することを試みる．また，様々なソーシャルネットワーク構造を定義し，それらの構造が検知に与える影響を横断的に比較する．次章では，ネットワーク構造からの分散表現獲得手法について説明する．

3. ネットワーク構造の表現学習

近年のネットワークデータの大規模化に伴い，より低次の空間でネットワークを表現する手法が盛んに開発されている．ネットワークの持つ情報を低次元で正確に表現することができれば，クラスタリングや分類問題，推薦システムなどの様々なタスクが適用可能となる．

LINE は，大規模ネットワークからの高速かつ正確な分散表現の獲得が可能な手法の 1 つである [Tang 15]．LINE には，ノード間の直接的な接続に基づく局所的な構造を保存する LINE(1st) と，ノードの共有に基づく大域的な構造を保存する LINE(2nd) の 2 種類の手法が提案されている．どちらの手法もエッジの重みを入力可能であるが，LINE(1st) は基本的に無向ネットワークのみを対象としているのに対して，LINE(2nd) は有向ネットワークにも適用可能である．本稿で取り扱うネットワークは全て重み付き有向ネットワークであるため，今回は LINE(2nd) のみを利用する．

LINE(2nd) の学習プロセスを説明する前に，本章で扱うネットワークを定義する．ノード集合 V，および有向エッジ集合 E が与えられたとき，有向ネットワークは $G = (V, E)$ と定義される．各エッジ $e \in E$ はノードの順序付きペア $e = (u, v)$ であり，ノード間の接続の強さを表す重み $w_{uv} > 0$ を持つ．LINE(2nd) の目的は，各ノード $v \in V$ を，ある低次元空間 R^d で表現することである．ただし，$d \ll |V|$ である．

LINE(2nd) では，あるノードは他のノードの文脈という役割が与えられる．そして，文脈にわたり類似の分布を有するノードは類似していると仮定する．LINE(2nd) は，この仮説を経験的に表現した確率分布 $\hat{p}(\cdot|v_i)$ と，分散表現ベクトルの内積により得られる確率分布 $p(\cdot|v_i)$ との間の差を最小化するように学習する．LINE(2nd) が解く目的関数は式 (1) で定式化される．

$$O = \sum_{i \in V} \lambda_i d\left(\hat{p}(\cdot|v_i), p(\cdot|v_i)\right) \tag{1}$$

ここで，$d(\cdot, \cdot)$ は 2 つの確率分布間の距離であり，相対エントロピーによって算出される．λ_i はネットワーク内のあるノード i の重要度を表しており，$\lambda_i = d_i$，$d_i = \sum_{k \in N(i)} w_{ik}$ である．ここで $N(i)$ はノード v_i の出次近傍である．

経験的確率分布は式 (2)，分散表現の内積から得られる確率分布は式 (3) で定義される．

$$\hat{p}\left(v_j|v_i\right) = \frac{w_{ij}}{d_i} \tag{2}$$

$$p\left(v_j|v_i\right) = \frac{\exp\left(\vec{u}_j'^T \cdot \vec{u}_i\right)}{\sum_{k=1}^{|V|} \exp\left(\vec{u}_k'^T \cdot \vec{u}_i\right)} \tag{3}$$

ここで，$\vec{u}$ はノードの役割の時の v_i の表現であり，$\vec{u}_i'$ は特定の文脈として扱われる時の v_i の分散表現を意味する．$|V|$ は文脈の数である．

LINE はまた，最適化プロセスにおいて，negative sampling およびエッジサンプリングにより，高速かつ正確な表現学習を実現している．最適化手法の詳細は [Tang 15] を参照されたい．

4. 計算機実験

以下ではまず，実験に使用するデータセットや事前パラメータ，評価方法について説明する．その後，実験結果について議論する．

4.1 cyber-predator の予測

4.1.1 データセット

本研究では，株式会社ナナメウエ [*1] が運営している「ひま部」[*2] という複数交流形 SNS において蓄積されたデータを使用する．ひま部は，メインターゲットを学生とした SNS であり，本研究のテーマに適したデータセットを保有している．ひま部が提供する主な機能は，昨今の主要な SNS と同様に，個人間や複数人でのチャット，文書や画像の投稿，そして投稿に対する返信や絵文字などによるリアクションである．ひま部では，サイバーリスクの自己対策のために，特定のユーザをブロックする機能や，ユーザや投稿などを運営に通報するといった行動が可能である．また，誘い出しや公序良俗に反する投稿などを行なったと運営が判断した場合，アカウントが停止され，今後一切サービスの利用ができなくなる．今回の実験では，アカウントが停止されたユーザを cyber-predator と定義する．

このようなサービス特性を踏まえ，我々は以下のアクションからソーシャルネットワークを定義する．

- フォロー：お気に入りのユーザを登録する行動．フォローを行うことにより，フォローしたユーザの投稿閲覧やチャットが容易になる．フォローしたユーザからフォローされたユーザに向かって有向エッジが接続される．フォローは 1 度しか行われないため，エッジの重みは 1 のみである．
- フォロー保留： フォローの申請があったにも関わらず，フォローを許可しない行動．これをネガティブな意思表示と捉え，フォローとは別のネットワークとして定義する．エッジの重みは 1 のみである．
- 足跡： 他ユーザのプロフィールページの閲覧行動．プロフィールを閲覧する行為は双方向的なコミュニケーショ

*1 https://nanameue.jp/ja
*2 https://himabu.com/

ンではないが，特定のユーザに対する何かしらの興味を表す行動であると考えられる．プロフィールを閲覧したユーザから閲覧されたユーザに向かって有向エッジが接続される．エッジの重みは閲覧回数とする．

- チャット：1 対 1 でのメッセージの送受信履歴．メッセージを送信したユーザから受信したユーザに向かって有向エッジが貼られる．エッジの重みは送信回数とする．
- コメント： 他ユーザの特定の投稿に対し，コメントする行動．投稿したユーザと，コメントしたユーザがエッジで結ばれる．エッジの重みはコメント回数とする．
- リアクション： 他ユーザの特定の投稿に対し，絵文字によりポジティブな感情を示す行動．リアクションしたユーザから投稿したユーザへ向かって有向エッジが接続される．エッジの重みはリアクション回数とする．
- ブロック： 特定のユーザからの個人チャットの受信や検索結果への表示を制限する行動．ネガティブな感情を表すネットワークである．ブロックしたユーザからブロックされたユーザに対して有向エッジが接続される．エッジの重みは全て 1 である．
- 通報： 特定のユーザを，不適切なユーザであると運営に報告する行動．これも直接的な交流ではないが，ネガティブな感情を表すネットワークを作成可能である．通報したユーザから通報されたユーザに対して有向エッジが接続される．エッジの重みは全て 1 である．

各ネットワークの基本的な統計量を表 1 に示す．それぞれ

表 1: 各ネットワークの統計量

ネットワーク	ノード数	エッジ数	平均入次数	平均出次数
フォロー	334,099	10,776,111	32.8	61.2
フォロー保留	119,776	295,895	20.3	2.7
足跡	496,393	28,429,788	60.7	92.8
チャット	146,137	1,595,801	11.3	12.3
コメント	73,163	348,562	5.2	6.1
リアクション	227,050	10,029,191	46.7	73.9
ブロック	103,422	341,814	3.6	9.1
通報	160,802	236,514	2.0	3.6

大きさは異なるが，1,000 万以上のエッジを持つ比較的大規模なネットワークも対象となっている．また，ノードの数に対する次数の数から，比較的疎なネットワークであることも確認できる．

4.1.2 評価方法

前節で定義した各ソーシャルネットワークから，LINE(2nd) によりそれぞれ分散表現を獲得する．ネットワークの作成に使用した期間は，2018 年 8 月 28 日から 2018 年 9 月 28 日までの 1ヶ月間である．各ネットワーク構造の有用性を比較するために，分散表現を入力特徴とした 2 クラス分類問題を設定する．対象となるユーザは，2018 年 9 月 21 日までにインストールし，最終ログイン日が 2018 年 8 月 28 日以降のユーザ，かつ 2018 年 9 月 28 日までにアカウントが停止していないユーザである．対象ユーザのうち，2018 年 9 月 29 日から 10 月 31 日の間にアカウント停止されたユーザを Predator クラス，それ以外のユーザを Normal クラスに分割する．

分類アルゴリズムには Random Forest [Liaw 02] を使用する．ただし，全体に占める Predator クラスの割合は 1%未満と極端に不均衡なデータセットであり，Random Forest を単に適用するだけでは良いモデルを構築することが困難である．また，各ネットワークは対象となるユーザが異なる．例えば，足跡やフォローといった行動は多くのユーザが行うが，通報を行うユーザは比較的少ない．そのため，クラス比にも違いが生じており，そのままでは正確な比較が難しい．

そこで我々は，アンダーサンプリング及び擬似データの生成によって，各ネットワークのクラス比を統一する．まず，ランダムなアンダーサンプリングにより，クラス比を 1:9 にする．その後，Synthetic Minority Over-sampling Technique (SMOTE) [Chawla 02] により，クラス比が 3:7 になるように，Predator クラスの擬似データを生成する．SMOTE は，元のデータに近い新たなインスタンスを生成することが可能であり，オーバーサンプリングによく利用される．モデルの評価には 10 分割交差検証法を利用し，F1 値および F1 値を構成する Precision と Recall により評価する．

4.1.3 事前パラメータ

LINE(2nd) には，いくつか事前に設定するパラメータが存在する．分散表現の次元数，negative sampling 数，そしてエッジサンプリング数である．次元数と negative sampling 数は，全てのネットワークでそれぞれ 100 と 5 に設定する．エッジサンプリング数は，各ネットワークの総エッジ数の約 50%に設定する．

Random Forest を構成する決定木の数は 50，分岐ルールには Gini 係数を使用する．各決定木の最大深さは 7，各葉が分岐を行うための最小インスタンス数は 20 に設定する．また，今回は不均衡データであるため，クラス比に応じてインスタンスの重みを調整する．

4.2 計算結果

以上の条件により，計算機実験を行なった結果を表 2 に記載する．全体的な評価をすると，対象とした 8 つのネットワー

表 2: 評価値

ネットワーク	F1 値	Precision	Recall
フォロー	0.571	0.620	0.530
フォロー保留	0.601	**0.670**	0.545
足跡	**0.637**	0.626	**0.648**
チャット	0.563	0.583	0.544
コメント	0.565	0.653	0.497
リアクション	0.505	0.522	0.502
ブロック	0.586	0.608	0.566
通報	0.560	0.631	0.504

クはある程度の説明力を有するが，極端に高い予測性能のモデルは存在しておらず，ネットワーク構造情報だけでは予測に限界があることが確認できる．

それぞれのネットワークの評価値を比較していく．最も F1 値が高かったネットワークは足跡であり，これは興味深い結果である．足跡は，フォローやチャットからなる一般的なネットワークに比べ，比較的弱い繋がりであるが，その分敷居の低い接触である．この結果から，足跡には，どのようなユーザに声を掛けるべきか，獲物を見定めているパターンが隠れているのではないかと考えられる．単純な足跡をつけた回数を比較しても，Normal クラスの中央値は 15 回であるのに対し，Predator クラスは 183 回であったことからも，Predator が多くのユーザのプロフィールにアクセスしていることが分かる．Predator クラスに顕著な足跡ネットワーク構造のパターンを明らかにすることは，今後の課題である．

足跡の次に F1 値が高いネットワークは，フォロー保留とブロックであった．これらは，ネガティブな感情を表すネットワークであるため，説明力が強いことは直感的にも理解できる．フォローとフォロー保留を分けてネットワークの構築を行うことが有用であることが明らかとなった．

逆に予測性能が相対的に低いネットワークは，リアクションと通報であった．リアクションは，足跡の次に大規模で緩いコミュニケーションであるが，Predator クラスの識別にはあまり寄与しないようである．一方の通報は，クラス定義と密接に関係しているように思われるが，今回の分類タスクが未然検知であり，既にアカウント停止されたユーザが対象外であったため，まだ通報という形で顕在化していない状態であることが原因であると考える．また，表 1 からも分かるように，通報ネットワークは特に疎なデータであった．[Tang 15] でも議論されているが，次数の少ないノードの分散表現を得ることは一般的に困難である．今後は，[Tang 15] で行われていた前処理のように，2 次近傍の接続まで考慮するなどの工夫により，ネットワークの密度も高くした上での比較実験も行なっていく．

5. おわりに

本稿では，ソーシャルネットワーク構造を利用した cyber-predator 未然予測モデルの構築を行なった．本稿の貢献は，これまであまり注目されていなかったプロフィール情報の閲覧や，フォローの保留など，様々なネットワークを定義し，公正な比較分析を行なった点である．モデルには，大規模かつ疎な重み付き有向グラフのネットワークに対して LINE(2nd) を適用し，構造を保存した分散表現を入力として与えることで，意味のあるモデルの構築に成功した．本稿で得られた知見は以下のとおりである．

- 多くのソーシャルネットワーク構造が，cyber-predator の検知に有効である
- 直接的な会話ネットワークよりも，プロフィール情報の閲覧という弱い繋がりが，cyber-predator の未然検知において強い説明力を持つ
- フォローとフォロー保留を区別することで，予測性能が向上する可能性がある
- ユーザのブロックや運営への通報により作成されたネットワーク構造は，未然検知においては説明力が比較的弱い

今後は，単に予測するだけではなく，cyber-predator 集合にはどのような構造パターンが出現しているのかを分析し，予測の妥当性を明らかにしていく．また，テキストコーパスやユーザ属性から得られる特徴も追加し，予測モデルのさらなる性能向上を図り，実サービスへの実装を通じてサイバー犯罪の抑止に貢献していく．

謝辞

本研究は RISTEX「未成年者のネットリスクを軽減する社会システムの構築」プロジェクトの助成を受けた研究である．また，貴重なデータをご提供いただいた株式会社ナナメウエの皆様に感謝申し上げます．

参考文献

[Cardei 17] Cardei, C. and Rebedea, T.: Detecting sexual predators in chats using behavioral features and imbalanced learning, *Natural Language Engineering*, Vol. 23, No. 4, pp. 589–616 (2017)

[Chawla 02] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique, *Journal of artificial intelligence research*, Vol. 16, pp. 321–357 (2002)

[F. Toriumi 15] F. Toriumi, M. T., T. Nakanishi and Eguchi, K.: Analysis of User Behavior on Private Chat System, in *2015 IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 1–4 (2015)

[Huang 14] Huang, V. K. S., Qianjia and Atrey, P. K.: Cyber Bullying Detection Using Social and Textual Analysis, in *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, pp. 3–6, New York, USA (2014), ACM

[Huang 18] Huang, V. K. S., Qianjia and Atrey, P. K.: On cyberbullying incidents and underlying online social relationships, *Journal of Computational Social Science*, Vol. 1, No. 2, pp. 241–260 (2018)

[Liaw 02] Liaw, A., Wiener, M., et al.: Classification and regression by randomForest, *R news*, Vol. 2, No. 3, pp. 18–22 (2002)

[Liu 17] Liu, D., Suen, C. Y., and Ormandjieva, O.: A Novel Way of Identifying Cyber Predators, *arXiv preprint arXiv:1712.03903* (2017)

[Nahar 13] Nahar, V., Li, X., and Pang, C.: An effective approach for cyberbullying detection, *Communications in Information Science and Management Engineering*, Vol. 3, No. 5, p. 238 (2013)

[Tang 15] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q.: LINE: Large-scale Information Network Embedding., in *WWW*ACM (2015)

[警察 17] 警察庁：平成 29 年上半期におけるコミュニティサイト等に起因する事犯の現状と対策 (2017)