ネットワーク分析を用いた誘い出し・ネットいじめの検知

東京大学　大学院工学系研究科
西口　真央

第２７回社会情報システム学シンポジウム　基調講演（オンライン） 2021.1.28

プロジェクトの全体像

未成年者が抱える
観測困難なネット利用リスクの軽減※

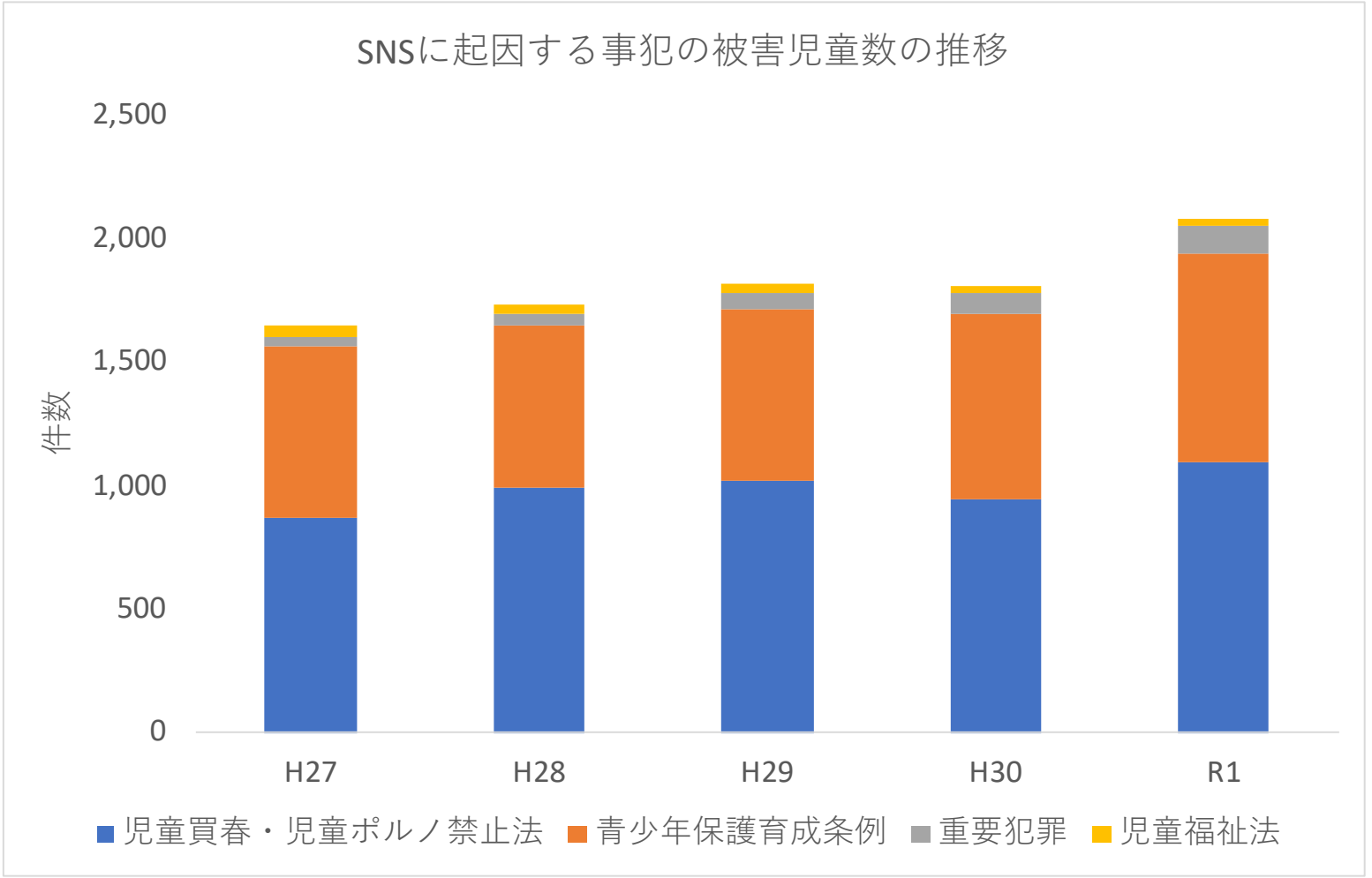
グループ	関係組織	テーマ
システム グループ	東京大学 サイバーエージェント ナナメウエ エースチャイルド LINE	未成年者のネットリスクを軽減するための **リスク検知システムの開発**
教育 グループ	関東学院大学 エースチャイルド	未成年者が自らリスクに気づくための **ネットリスク教育法の開発**
社会 グループ	中央大学	ステークホルダーにシステム導入を促すための **インセンティブ設計**

※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, https://www.jst.go.jp/ristex/pp/project/h29_3.html

※「少女たちの10日間」オフィシャルサイト, http://www.hark3.com/sns-10days/

誘い出し等の被害の現状

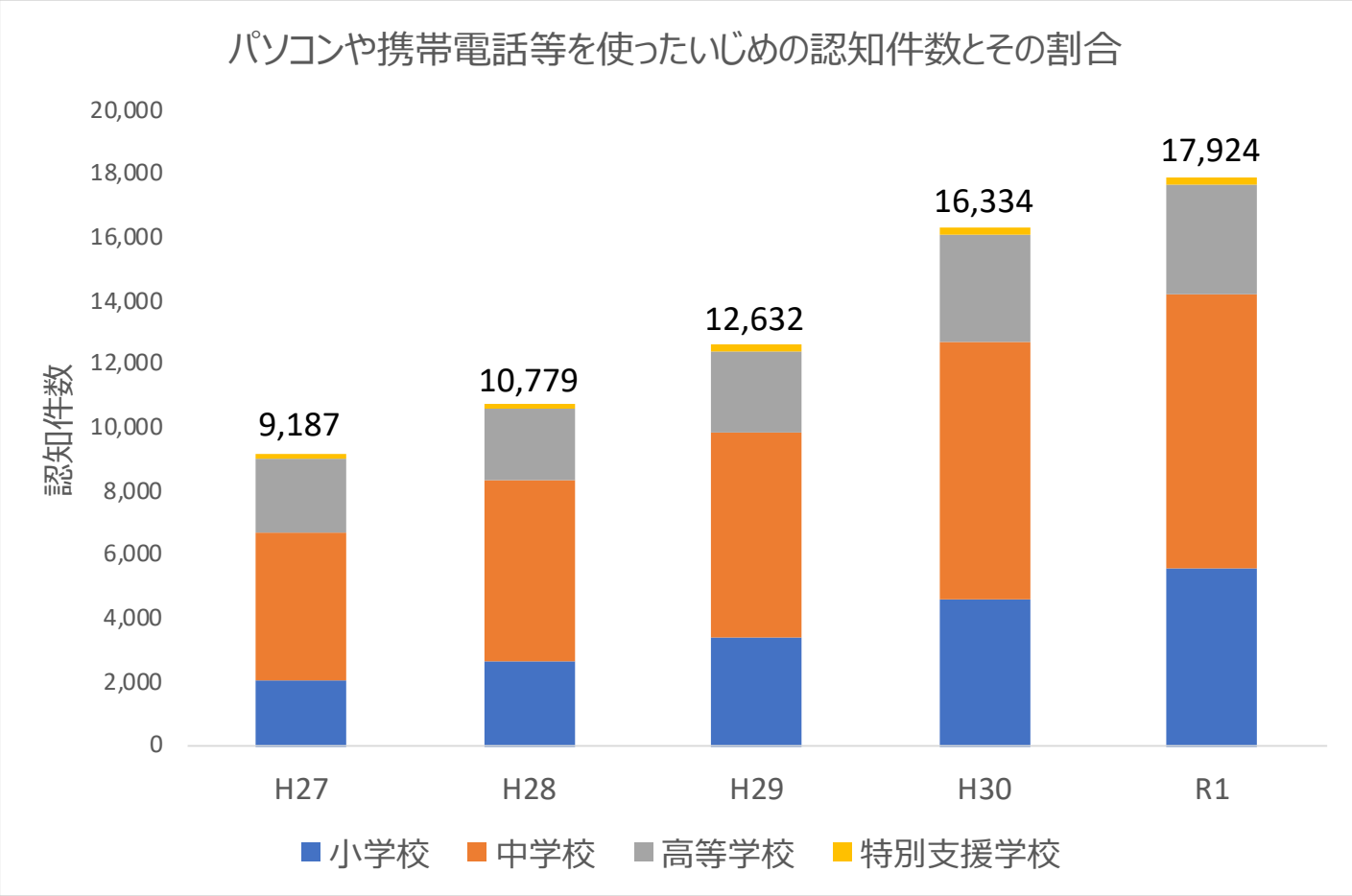
■ 全体の件数、重要犯罪の件数が増加傾向
- 座間事件、茨城女子大生殺人事件、某youtuberなど



引用：警察庁「令和元年における少年非行、児童虐待及び子供の性被害の状況」, https://www.npa.go.jp/safetylife/syonen/hikou_gyakutai_sakusyu/R1.pdf

ネットいじめ被害の現状

■ 同様に増加傾向
- 特に小学校の認知件数が顕著



引用：文科省「令和元年度児童生徒の問題行動・不登校等生徒指導上の諸課題に関する調査」, https://www.mext.go.jp/a_menu/shotou/seitoshidou/1302902.htm

実装する２つのシステム

- 「誘い出し加害者」と「いじめ被害者」の検知システムを構築
- 実サービスでの運用を見据えた共同研究

誘い出しユーザの検知

- ピグパーティとの共同研究
- 誘い出し行為を行なっている可能性の高いユーザ集合を出力
- 有人監視を経て、ペナルティの付与やアカウントBANの実行

いじめ被害者の検知

- 子供見守りアプリfiliiとの共同研究
- ネットいじめの予兆を確率で出力
- 本人と保護者に自動アラート送信，啓蒙情報の提供

誘い出し
ユーザの検知

ピグパーティとは

仮想空間内でなりきりたいアバター（ピグ）を作って、
ピグのきせかえや、自分のお部屋のもようがえをしながら楽しむ
アバターコミュニケーションアプリです。

GET IT ON Google Play

Download on the App Store

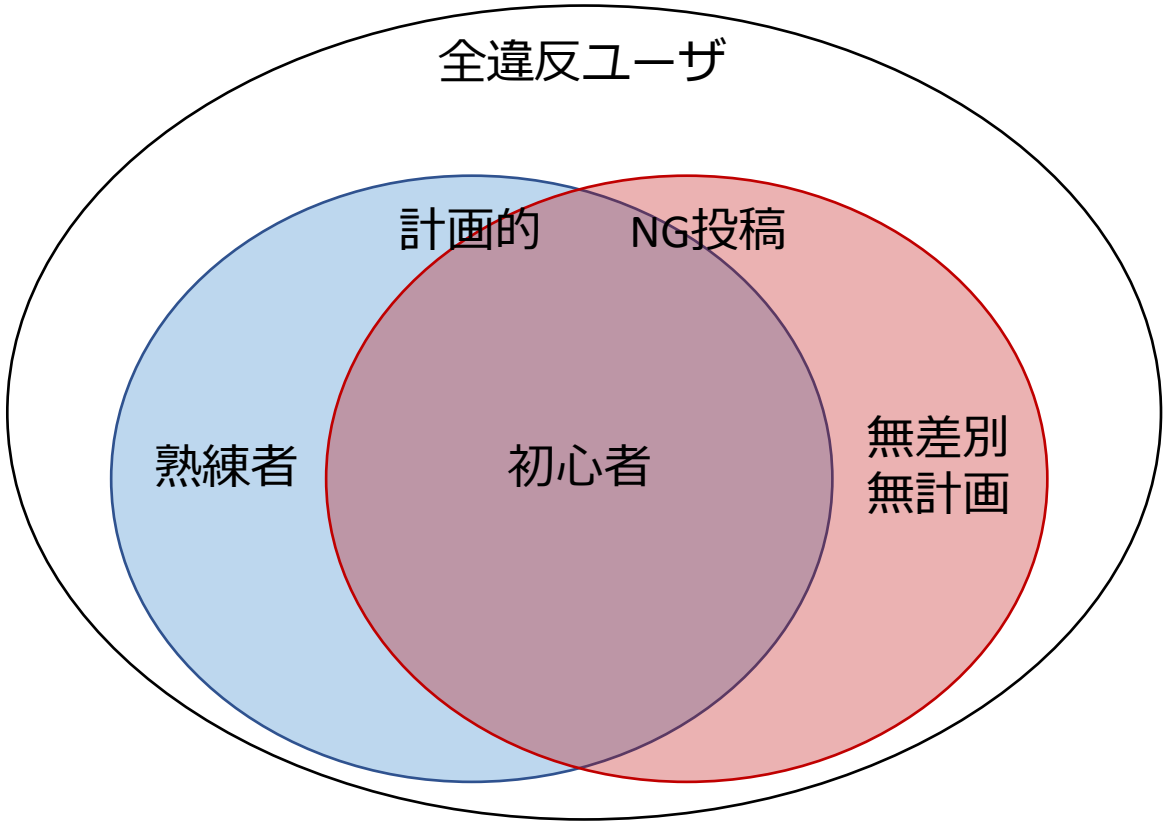
違反ユーザのカテゴライズ

■ 計画的 or 無差別

- 無差別な違反は低レベルなものが多い
- まず検知すべきは計画的な違反ユーザ

■ 初心者 or 熟練者

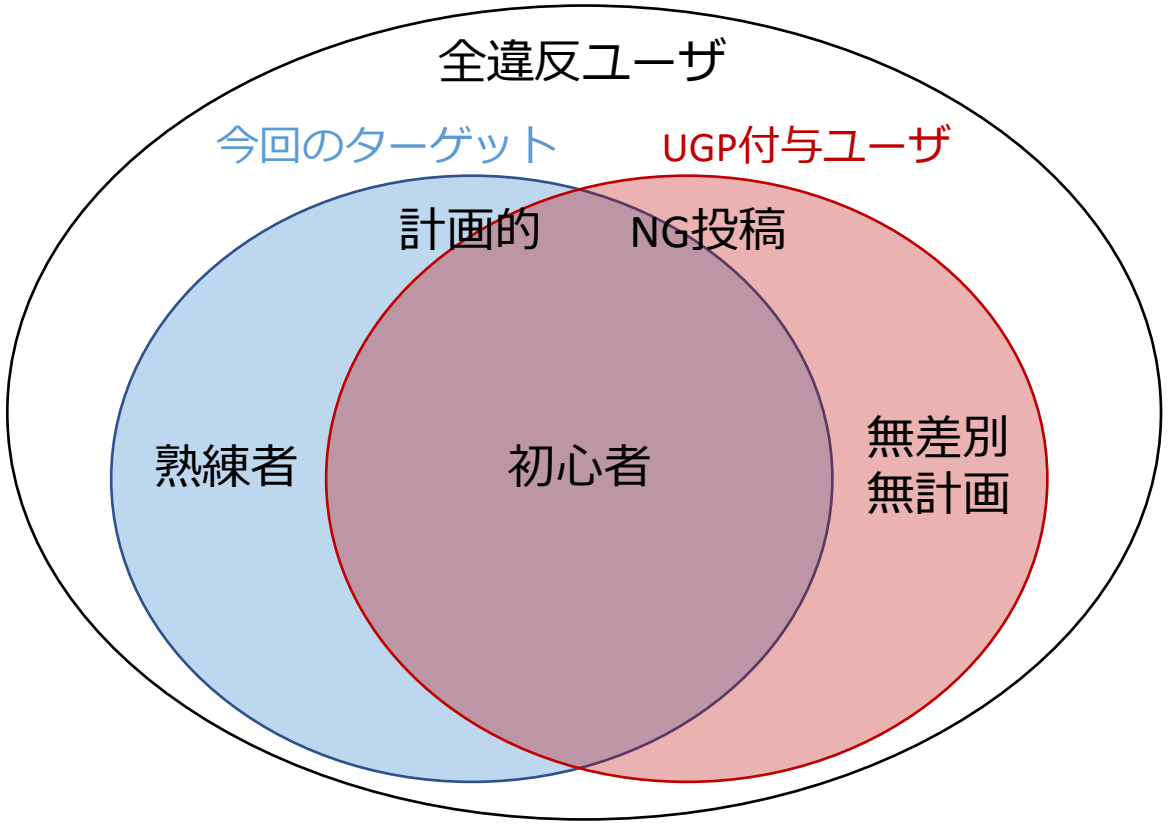
- 熟練者をNGワードで検知することは困難



本システムの役割

- たとえNGワードを学習しても、接触するユーザは変わらないのでは？

- **計画的な初心者の接触行動をモデル化して、熟練者を発見**



SNS違反検知システムの重要ポイント

■ プライバシー

- 通信の秘密：プライベートなテキストデータの利用が困難
- 個人情報保護：ユーザを特定し得るデータの利用が困難

■ 全自動化が困難

- AIの暴走
- 「怪しい」だけではBANできない

■ 多様な機能

- トーク、パーティ、コミュなど多様な機能があり、SNSによっても異なる

■ コンテンツの流動性

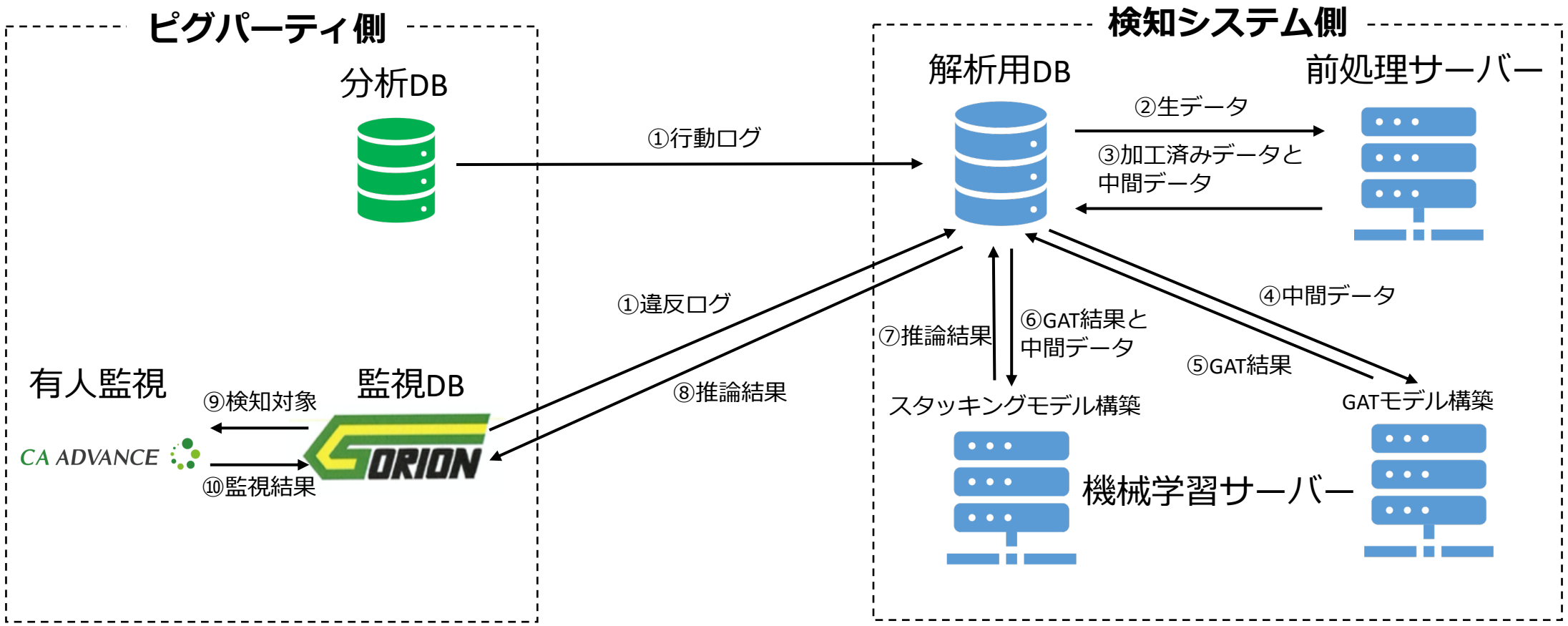
- チャットルームや投稿などのコンテンツは常に入れ替わる

■ クラスの不均衡性

- 違反ユーザの割合は1%以下と極端に不均衡

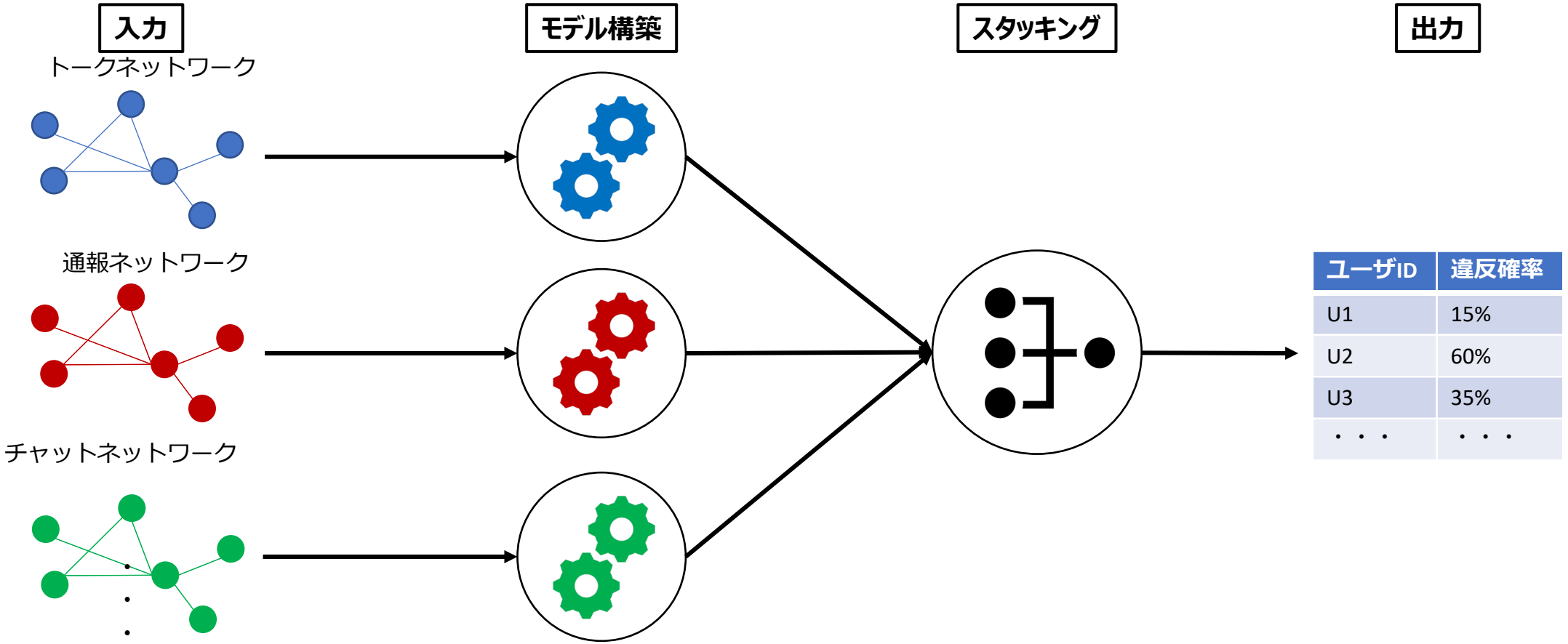
システムの仕組み（データの流れと役割）

- 違反ログと、ユーザIDが暗号化された**行動ログ（テキスト以外）**を受け取る
- 検知システムは推論結果を返す
- **全自動ではなく、有人監視による最終ジャッジ**



モデル構築の流れ

- **様々なユーザ関係ネットワーク**を利用した、クラス分類問題による検知
- 個人間メッセージ含め、個人情報を一切必要としない
- ネットワークごとにモデル構築し、スタッキングにより統合利用

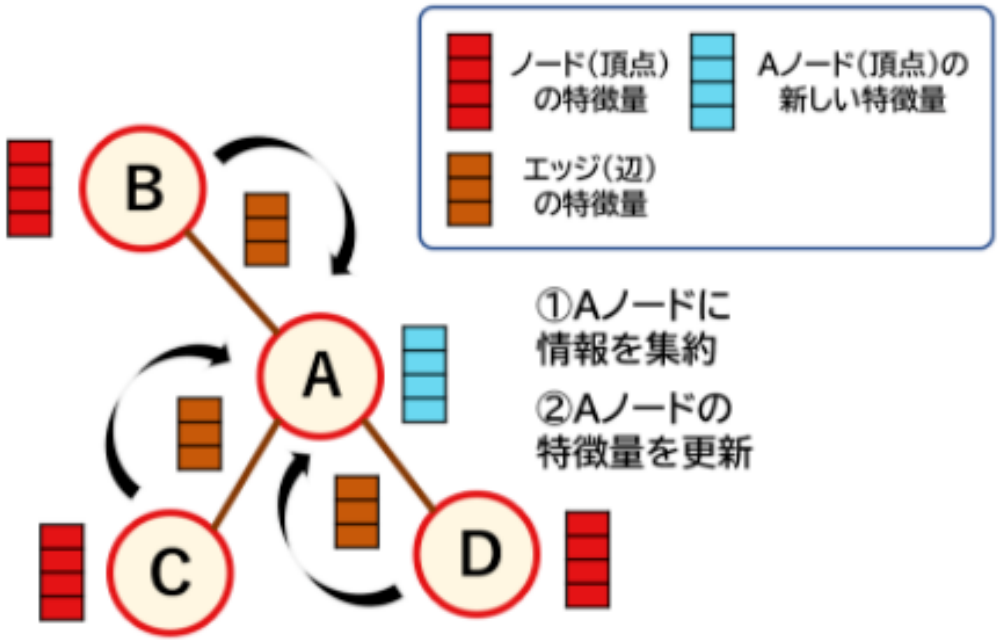


ユーザID	違反確率
U1	15%
U2	60%
U3	35%
・・・	・・・

モデル構築手法：Graph Attention Networks (GAT)

- ネットワークデータに対するディープラーニング系手法の１つ
  - ノード：ユーザの特徴量（性別、年代、インストール日など）
  - エッジ：ユーザ同士の何かしらの繋がり（フレンド関係、ブロック関係など）
- 自身のノードの持つ特徴量と、直接繋がりのある近接ノードの特徴量を考慮（畳み込み演算）
- 繋がりの重要性をエッジの重みとして学習（**アテンション**メカニズム）

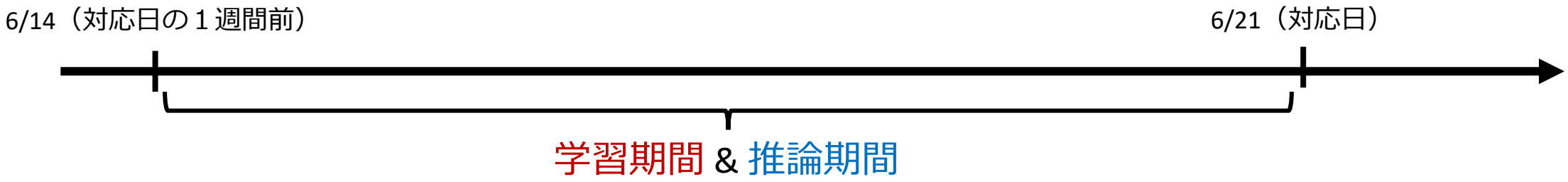
図. ユーザAの学習過程のイメージ



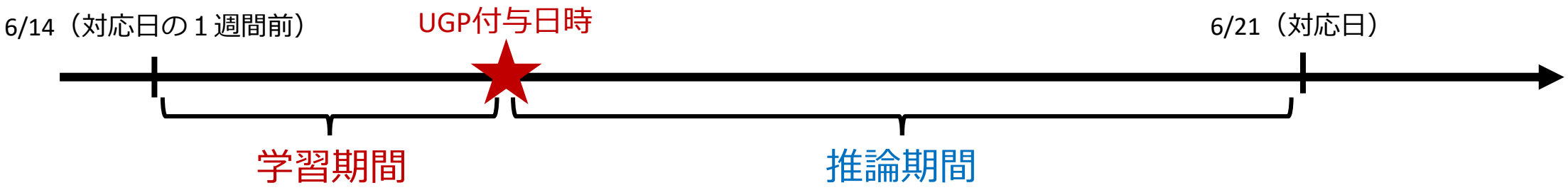
設定タスク　学習期間

- 直前 1 週間のデータを利用して学習 & 推論
- 違反履歴のあるユーザは、違反前と違反後で別のユーザとして扱う
- 違反前のデータは推論には利用しない

【直前 1 週間で**違反のない**ユーザ】



【直前 1 週間で**違反した**ユーザ】



利用するデータ：ユーザ関係ネットワーク

- 6つの行動タイプからネットワークを作成
- コンテンツの流動性を考慮し、全てユーザ関係ネットワークとして定義

行動タイプ	ネットワーク
トーク	送信ユーザ => （トークルーム） => 受信ユーザ
チャット	発話ユーザ => （パーティルーム） => オーナーユーザ
	発話ユーザ => （お部屋） => オーナーユーザ
フレンドリクエスト	申請したユーザ => 申請されたユーザ
フォロー	フォローしたユーザ => フォローされたユーザ
コメント	コメントしたユーザ => （投稿） => コメントされたユーザ
いいね	いいねしたユーザ => いいねされたユーザ

利用するデータ：ユーザ（ノード）の特徴量

特徴量	要素
年代	不明、１０歳以下、１２歳以下、１５歳以下、１８歳以下、２２歳以下、３０歳以下、５０歳以下、それ以上でカテゴライズ
性別	不明、男、女
フレンド数	予測実行時点のトータルフレンド数
インストール後経過日数	０日、３日以内、７日以内、１４日以内、それ以上でカテゴライズ
各アクションの有無	• 入場した回数（エリア、お部屋、パーティ、その他を区別） • チャット投稿数（エリア、お部屋、パーティ、その他を区別） • 入場した場所数（エリア、お部屋、パーティ、その他を区別） • チャット投稿した場所数（エリア、お部屋、パーティ、その他を区別） • フィード投稿数 • フィード投稿に対するコメント数 • フィード投稿へのいいね数 • フォロー数 • フレンド申請数 • ギフト送信数 • パーティ作成数 • プロフィール閲覧数 • トークへの投稿数（グループ、プライベートを区別） • トークUU数（グループ、プライベートを区別） • きゅん数 • お部屋デコ回数 • 着せ替え回数

対象ユーザ

【対象ユーザ】

- 直前７日間で下記の条件のいずれかを満たすユーザを対象に、それぞれ学習用データセットを作成

トークユーザ数	３以上
チャットしたユーザ数	
フレンド申請数	
フォロー申請数	
コメントしたユーザ数	
いいねしたユーザ数	20以上

- 推論対象ユーザは、直前７日間にログインした全ユーザ（永久BANユーザを除く）

【設定クラス】

- 違反クラス：対象ユーザのうち、直前１週間で一時BANを受けたユーザ
- 健全クラス：対象ユーザのうち、直前３０日間でペナルティを1度も受けていないユーザ

入力データの基本統計量

- 各ネットワークで分布も対象ユーザも異なる
- 比較的大規模で疎なネットワーク
- クラスのないノードも一定数存在
- 不均衡分類問題

ネットワークタイプ	ノード数	エッジ数	1 人あたりエッジ数	違反クラスノード数	健全クラスノード数	クラスなしノード数	違反クラスの割合
リクエスト	46,892	486,258	10.3	244	37,169	9,479	0.7%
チャット	51,151	1,247,344	24.3	344	38,165	12,642	0.9%
トーク	26,764	293,390	10.9	235	18,326	8,203	1.3%
コメント	16,702	163,638	9.7	137	11,544	5,021	1.2%
フォロー	6,324	155,576	24.6	54	4,338	1,932	1.2%
いいね	81,040	10,228,766	126.2	334	66,942	13,764	0.5%

実験設定　モデル全般

【予測実行日】

- 予測対象日：２０２０年１０月１４日
- 学習データ期間：２０２０年１０月８日～１４日

【テストデータの評価（初心者の行動をうまく学習できたか）】

- 訓練データ７割（うち検証データ３割）、テストデータ３割
- AUCを最も重視して評価
- 予測確率50%以上とした場合の、Accuracy、F1、Precision、Recallも算出
- スタッキング前の、各GATモデルも評価

【人力評価（熟練者を発見できたか）】

- 予測確率が上位のユーザと、ランダムに抽出されたユーザ約2,000人を、有人監視チームが熟練違反者か否かラベル付け
- 評価結果から、適切な予測確率のしきい値を決定

実験設定　GAT

- 損失関数：交差エントロピー誤差
- 活性化関数：LeakyReLU
- 最適化関数：Adam SGD
- エポック数：300
- 早期停止のチェック間隔：100
- マルチヘッド数：8
- 隠れ層の数：1
- 隠れ層の次元数：8
- 学習時には、クラス比率に応じたクラスの重みを設定

- 以下のパラメータをグリッドサーチして、検証データの損失が最小となるモデルを選択

パラメータ	値
学習率	0.006～0.01（0.001刻み）
特徴量のドロップアウト率	0.1～0.3（0.1刻み）
アテンションのドロップアウト率	0.1～0.3（0.1刻み）
重み減衰の係数	{0.01～0.014}（0.002刻み）
ReLUの負の勾配	{0.1～0.3}（0.1刻み）

実験設定　LightGBM

- スタッキングにはLightGBMを利用
  - 特徴量の欠損値をそのまま利用可能
- 訓練データは、クラス比率が95:5となるように健全クラスをランダムアンダーサンプリング
- さらに、クラス比率が90:10となるように違反クラスをSMOTEオーバーサンプリング
- 学習時には、クラス比率に応じたクラスの重みを設定
- 以下のパラメータをグリッドサーチして、検証データのAUCが最小となるモデルを選択

パラメータ	値
基本学習器の数	5～51（5刻み）
特徴学習率	0.025～0.05（0.005刻み）
葉ノードの数	2～10（2刻み）
１つの基本学習器の最大深さ	2～10（2刻み）
１つの葉ノードに必要な最小サンプル数	10～30（10刻み）

実験結果（テストデータによる評価）

- スタッキング（ALL）によるAUCが最も高く、0.932と高性能
- 違反ユーザは熟練者である可能性もあるので、テスト結果としては十分

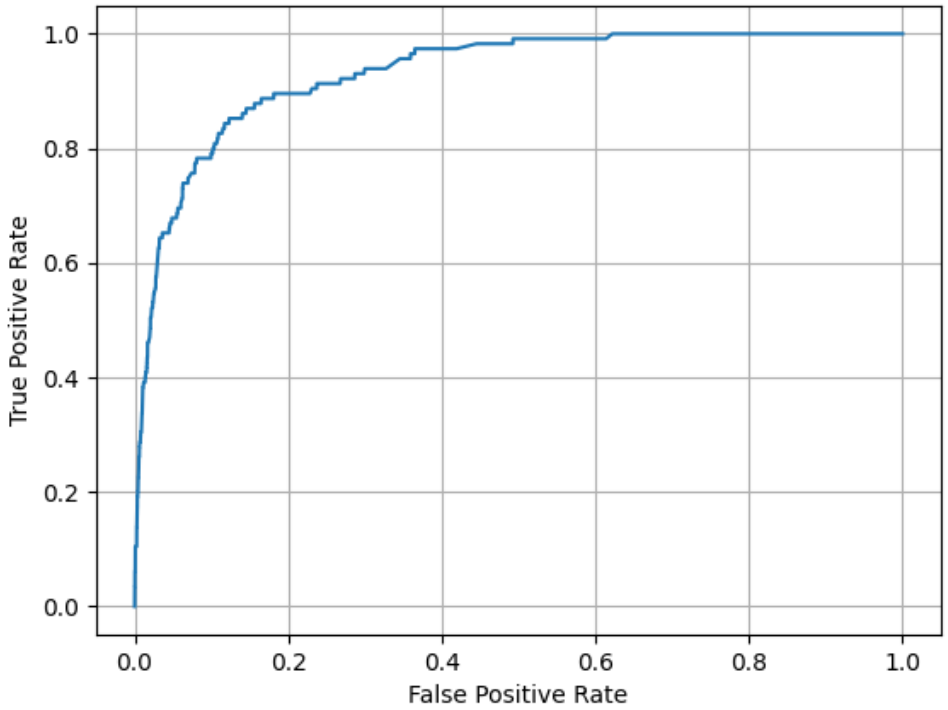
評価値

モデル	Accuracy	Precision（違反）	Recall（違反）	F1（違反）	AUC
リクエスト	0.827	0.021	0.575	0.041	0.790
チャット	0.830	0.032	0.612	0.060	0.790
トーク	0.836	0.040	0.521	0.075	0.754
コメント	0.801	0.034	0.585	0.064	0.769
フォロー	**0.926**	**0.045**	0.250	**0.076**	0.690
いいね	0.721	0.014	0.790	0.027	0.825
ALL	0.876	0.033	**0.852**	0.063	**0.932**

ALLの分類表（>=50%）

		予測		
		健全	違反	合計
実測	健全	20,310	2,885	23,195
	違反	17	98	115
	合計	20,327	2,983	23,310

ALLのROC曲線



人力評価　監視の着目ポイントの絞り込み

■ １週間の全ての投稿やDMを有人監視するのはコストが高い

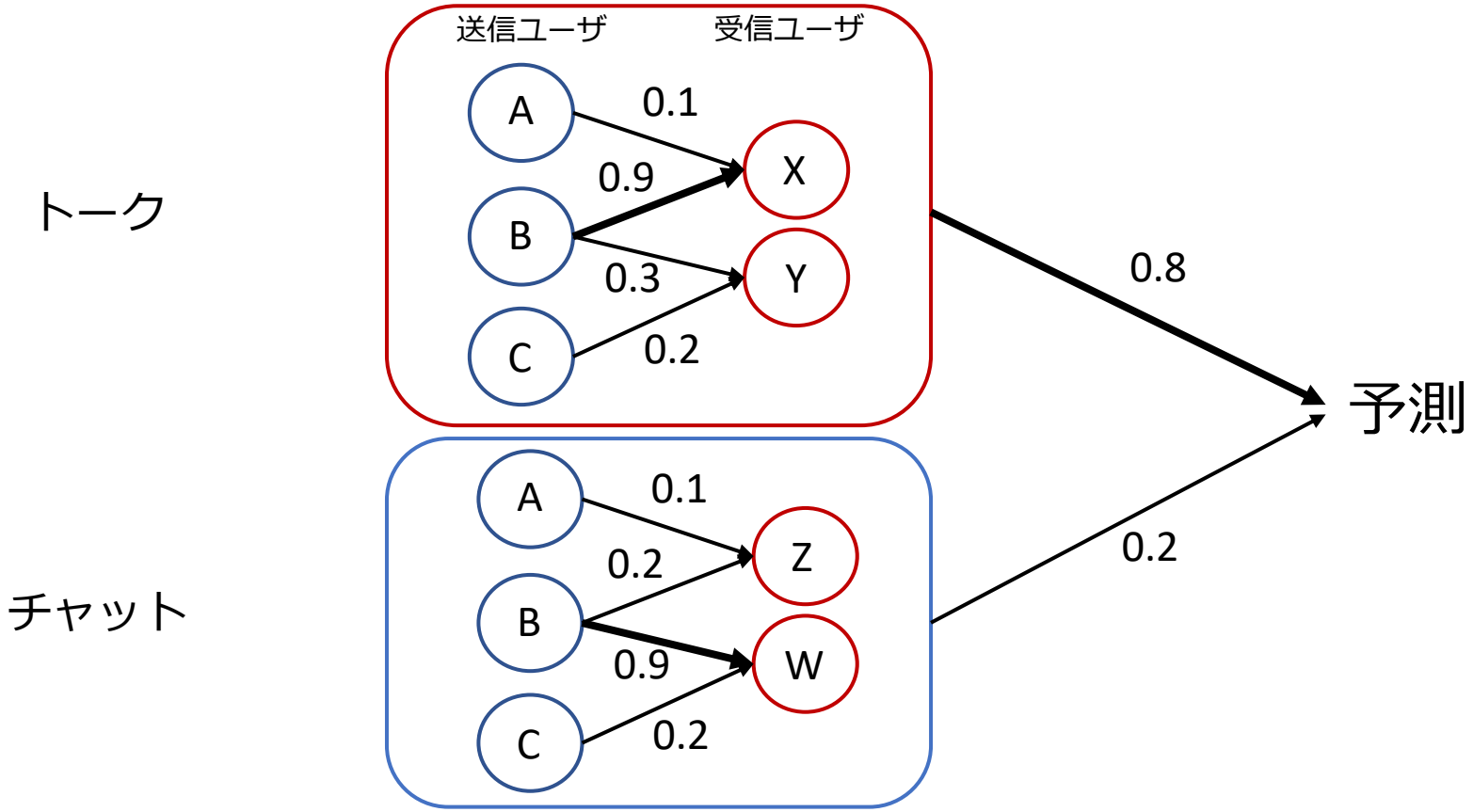
■ 「場所」と「相手」、「日付」の３つの観点で監視ポイントの絞り込みを行う

ユーザID（暗号化済み）	予測確率	着目ポイント
u_0001	98.8%	{‘place’: ‘talk’, ‘id’: [57947691, 8239123, ・・・], ‘date’: 2020/10/13}
u_0002	95.1%	{‘place’: ‘room’, ‘id’: [3659822, 37529123, ・・・], ‘date’: 2020/10/13}
u_0003	80.2%	{‘place’: ‘party’, ‘id’: [2480173, 8239123, ・・・], ‘date’: 2020/10/13}
u_0004	70.7%	{‘place’: ‘area’, ‘id’: [57947691, 8239123, ・・・], ‘date’: 2020/10/13}
u_0005	70.5%	{‘place’: ‘talk’, ‘id’: [57947691, 8239123, ・・・], ‘date’: 2020/10/13}

表. システムが返す結果例

監視の着目ポイントの絞り込み

- アテンション係数　　：相手の絞りこみに利用
- 各GATモデルの予測確率　：場所の絞り込みに利用
- ログイン履歴　　：時間の絞り込みに利用



まとめと今後

【まとめ】

- SNS上の「誘い出し加害者」と「いじめ被害者」の２つの汎用検知システムを構築
- 様々な問題をクリアした現実的な仕組みを確立
  - 主にユーザ関係ネットワークを利用することで、**プライバシー**に配慮
  - 有人監視を加えることで、**AIの限界**に配慮
  - スタッキングにより、**SNSの多様性**を考慮
  - ユーザ関係ネットワークに変換することで、**コンテンツ流動性**を緩和
  - オーバー/アンダーサンプリングとクラス重みによる**不均衡データ**への対応
- 本番運用まで見据えた共同研究による迅速な社会実装

【今後】

- 本番運用を行うことで、現場の声を取り入れた継続的な改善
- 事業化し、様々なSNS、ライブ配信アプリ、およびマッチングアプリに横展開