Active Link Inference for Case Building Investigation

Reihaneh Rabbany & David Bayani & Artur W. Dubrawski

Auton Lab
Carnegie Mellon University

How can we infer connections between entities, given the evidences of their relations? How can we help an investigator to find salient patterns by ***connecting the dots*** *more efficiently?*

> inferring links where the user is able to actively provide feedback and guide the inference

Online Human Trafficing

- ILO: 4.5 million people victimized globally by forced sexual exploitation in 2014; 21% were children
- MCMEC: 1,432% increase in reports of suspected child sex trafficking in 2014
- *correlated with the increased use of online sex advertisements*

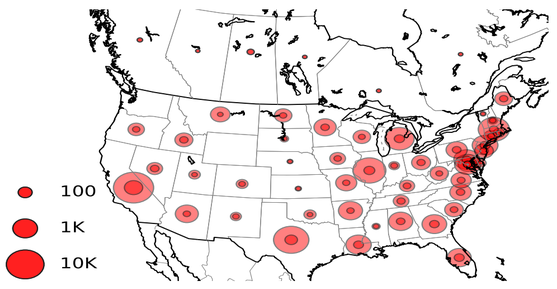

Figure 1: A wide-spread issue: reported human trafficking cases. Sample distribution of advertisements posted on Backpage.com, which contain a phone number reported to a human trafficking hotline.

Active Link inference for case building to combat human trafficking in online escort advertisements

find how different advertisements are linked together and point to organized activities; e.g. advertisements for different potential victims which are linked by phone numbers, catch phrases or text patterns, images with the same background, or other evidences of connection.

> an initial lead is treated as the seed query to find connected entities and identify the person of interest.

Data Source

Millions of escort advertisements scraped from Backpage.com for cities across the US and Canada posted between 2013 to 2017. For each advertisement, we have its unstructured text (title and body), attached images, date and location posted.

Problem Definition

- $\mathcal{D} = \{d_1, d_2 \dots d_n\}$ connected to k different types of evidences (e.g. phone number, image, bi-grams)
- $\mathcal{E} = \{\mathbf{X}_1, \mathbf{X}_2 \dots \mathbf{X}_k\}$ denote the set of indicator matrices for these k modalities i.e. $\mathbf{X}_m \in \mathbb{R}_+^{n \times c_m}$ for $m \in [1 \dots k]$
  - c_m is the cardinality (number of unique evidences) of modality m (e.g. number of unique phone numbers)

Each column of $\mathbf{X}_m$ shows a set of datapoints that share the corresponding evidence (e.g. datapoints that all share a particular phone number), and each row of $\mathbf{X}_m$ shows the evidences associated to the corresponding datapoint (all the phone numbers mentioned in a particular datapoint).

Active Link Inference *Given seed datapoint of interest, i, and assuming an unknown label set $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ where $y_j = 1$ if the user deems node y_j related to the node i and zero otherwise, find the maximum number of related entities to i given a fixed query budget, $b > 0$.*

Data Representation

Each indicator matrix: biadjacency matrix of a 2-partite graph > k-partite graph representation for the data

Given which and seed node of interest v_i, the Active Link Inference translates to finding positive nodes that are (tightly) connected to node v_i with length even paths (i.e. through shared evidences); whereas positive nodes means when we query label of the found endpoint v_j, we have $y_j > 0$.

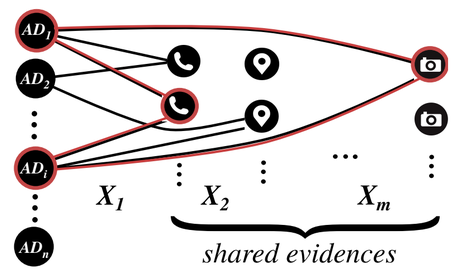

Figure 2: k-modal evidence graph: ads and evidences associated with them.

Our proposed method navigates through this graph to efficiently find related nodes while learning the importance of each modality and each piece of evidence from the labels (user's feedbacks) obtained while expanding.

Method Description

- $\mathbf{\Theta} = [\theta_1, \theta_2, \dots \theta_k]$; θ_m denotes the importance of modality m
- $\mathbf{s}^j = [s_1^j, s_2^j, \dots s_k^j]$; s_m^j shows the tie strength of reaching v_j from evidences in modality m
- $s_m^j = \sum_{v_i \in L^+} \sum_{u \in \mathcal{N}(v_i) \cap \mathcal{N}(v_j)} \frac{1}{d_u^2} - \sum_{v_i \in L^-} \sum_{u \in \mathcal{N}(v_i) \cap \mathcal{N}(v_j)} \frac{1}{d_u^2}$; where $L^+ = \{j \in L |\ y_j > 0\}$, and $L^- = \{j \in L |\ y_j < 0\}$
- **infer the next node**: $j^* \leftarrow \arg\max_j \sum_m s_m^j \theta_m^j$ the node with highest overall evidence support.
- **adjust the importance of responsible modality** :
- $m^* = \arg\max_m\ s_m^j \theta_m$
- $\theta_m^* = \delta \theta_m^*$ if $y_j < 0$ and $\theta_m^* = (2 - \delta)\theta_m^*$ if $y_j > 0$;
- $\delta \in (0, 1)$ is the learning rate

One Discovred Case

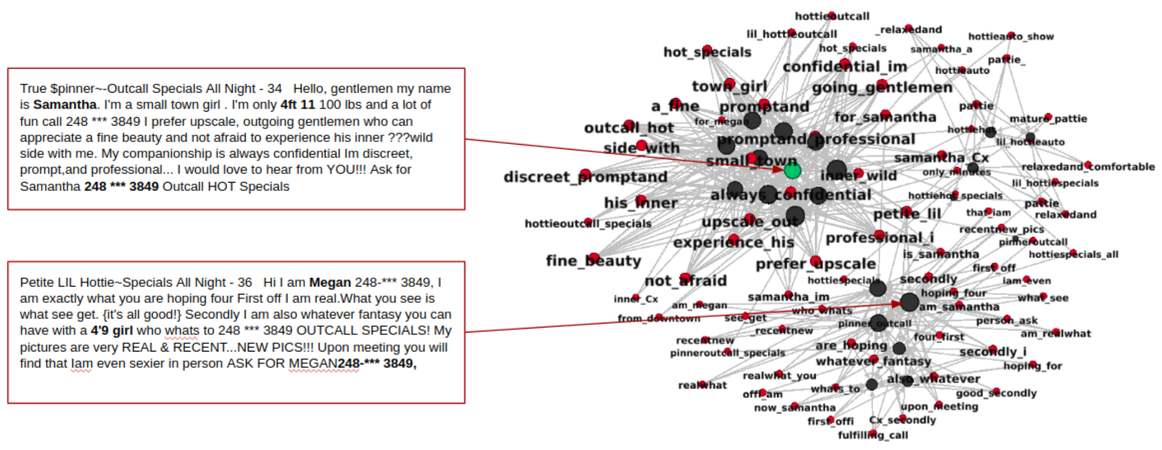

Figure 3: Green advertisement shows the seed advertisement used in this case study. The black nodes are the related advertisements discovered which are connected to the seed though shared evidences, plotted as red nodes. Although these two text don't show high similarity, these two ads are truly related as they share the same phone number.Out method was able to correctly discover it, and this representation explains how these ads are connected through shared evidences.

Performance Evaluation

- **No Feedback (NF)**: max score
- **No Modality (NM)**: re-weigh the evidences
- **No Evidence Weights (NE)**: learn the importance of modalities
- **Random Walk (RW)**: restarts from the seed whenever it reaches a negative node.

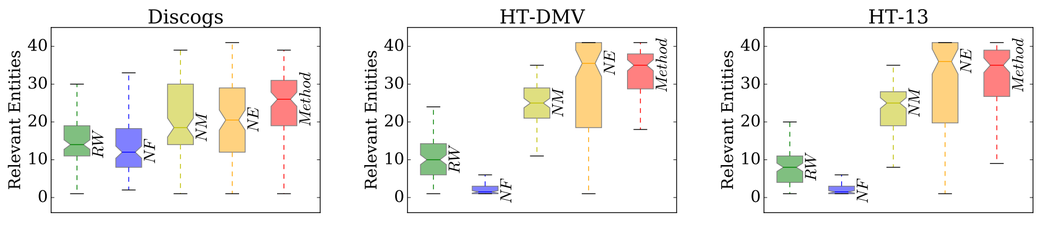

Figure 4: Comparison with baselines on three datasets: 2.5 million ads posted in DC, Maryland, Virginia area (HT-DMV), and another one of about 4 million ads posted between July to December 2013 (HT-13). Strong indicators (url, email, and phone numbers) as labels; as well as Discogs, a public online music database of 3.5 million different releases: date of the release, artists (primary and extra) involved, record labels, track information, companies involved in the production of the release, etc. Here, we find releases of the same artist as the given seed release based on their shared information.