Ranking in Heterogeneous Networks with Geo-Location Information

Abhinav Mishra
Amazon

Leman Akoglu
CMU
Ranking in networks

- Which nodes are the most important, central, authoritative, etc.?
  - Pagerank [Brin&Page, ‘98]
  - HITS [Kleinberg, ’99]
  - Objectrank [Balmin+, ’04]
  - Poprank [Nie+, ’05]
  - Rankclus [Sun+, ’09]
  - …
Ranking in rich networks

- How to rank nodes in a directed, weighted graph with multiple node types and location information?

- Different types of nodes ranked separately
Example

Weighted medical referral network (directed)
Example

Weighted medical referral network (directed) + physician expertise
Weighted medical referral network (directed) + physician expertise + location (distance)
Example

Ranking Problem: Which are the top $k$ nodes of a certain type?

e.g.: Who are the best cardiologists in the network, in my town, etc.?
Goal: ranking in directed heterogeneous information networks (HIN) with geo-location

- HINside model
- Parameter estimation
  - via learning to rank
- Experiments
Outline

**Goal**: ranking in directed heterogeneous information networks (HIN) with geo-location

**HINside model**

1. Relation strength
2. Relation distance
3. Neighbor authority
4. Authority transfer rates
5. Competition
   - Closed form solution

- Parameter estimation
- Experiments
HINside model

- Relation Strength and Distance
  - edge weights
    \[ W(i, j) = \log(w(i, j) + 1) \]
  - pair-wise distances
    \[ D(i, j) = \log(d(l_i, l_j) + 1) \]

(3.1) \[ M = W \odot D \]
HINside model

- In-neighbor authority

\[
(3.2) \quad r_i = \sum_{j \in \mathcal{V}} M(j, i) \ r_j
\]

\( r_i \): authority score of node \( i \)

- Authority Transfer Rates (ATR)

\[
(3.3) \quad r_i = \sum_{j \in \mathcal{V}} \Gamma(t_j, t_i) \ M(j, i) \ r_j
\]

\( t_i \): type of node \( i \)
HINside model

- **Competition**

\[ N(u, v) = \begin{cases} 
  g(d(l_u, l_v)) & u, v \in V, u \neq v \\
  0 & u = v 
\end{cases} \]

for monotonically decreasing \( g(z) = e^{-z} \)

\[(3.4) \quad r_i = \sum_j \Gamma(t_j, t_i) M(j, i) (r_j + \sum_{v: t_v = t_i} N(v, j) r_v) \]
Closed-form solution

- Authority scores vector $\mathbf{r}$ written in closed form as (& computed by power iterations)

$$
\mathbf{r} = \left[ L' + (L'N' \odot E) \right] \mathbf{r} = \mathbf{H} \mathbf{r}
$$

- $L = M \odot (T \Gamma T')$
- $T$ (n x m) where $T(i, c) = 1$ if $t_i = T(c)$
- $\Gamma$ (m x m) authority transfer rates (ATR)
- where $E(u, v) = \begin{cases} 
1 & \text{if } t_u = t_v \\
0 & \text{otherwise}
\end{cases}$

$$
E = TT'
$$

n: #nodes  m: #types
Outline

Goal: ranking in directed heterogeneous information networks (HIN) with geo-location

- HINside model
- Parameter estimation
  - via learning-to-rank objectives
- Experiments
Parameter estimation

- HINside’s parameters consist of the $m^2$ authority transfer rates (ATR)

\[(3.4) \quad r_i = \sum_j \Gamma(t_j, t_i) M(j, i) \left( r_j + \sum_{v: t_v = t_i} N(v, j) r_v \right)\]

- $r_i$ as a vector-vector product

\[
\begin{align*}
    r_i &= \sum_t \Gamma(t, t_i) \left[ \sum_{j: t_j = t} M(j, i) \left( r_j + \sum_{v: t_v = t_i} N(v, j) r_v \right) \right] \\
    &= \sum_t \Gamma(t, t_i) X(t, i) \\
    &= \Gamma'(t_i, :) \cdot X(:, i) = \Gamma'_{t_i} \cdot x_i \\
    &= f(x_i) = \langle w, x_i \rangle
\end{align*}
\]
An alternating optimization scheme:

$$\begin{align*}
\Gamma & \rightarrow r \rightarrow X^{\text{estimate}} \rightarrow \Gamma
\end{align*}$$

**Given:** graph $G$, (partial) lists ranking a subset of nodes of a certain type

- Randomly initialize $\Gamma^0, k = 0$
- Compute authority scores $r$ using $\Gamma^0$
- **Repeat**
  - $X^k \leftarrow$ compute feature vectors using $r$
  - $\Gamma^{k+1} \leftarrow$ learn new parameters by learning-to-rank
  - compute authority scores $r$ using $\Gamma^{k+1}$
- **Until** convergence
An alternating optimization scheme:

\[ \begin{align*}
\Gamma & \rightarrow r \\
& \rightarrow X \xrightarrow{\text{estimate}} \Gamma
\end{align*} \]

**Given:** graph \( G \), (partial) lists ranking a subset of nodes of a certain type

- Randomly initialize \( \Gamma^0, k = 0 \)
- Compute authority scores \( r \) using \( \Gamma^0 \)
- **Repeat**
  - \( X^k \) \( \leftarrow \) compute feature vectors using \( r \)
  - \( \Gamma^{k+1} \) \( \leftarrow \) learn new parameters by learning-to-rank
  - compute authority scores \( r \) using \( \Gamma^{k+1} \)
- **Until** convergence
RankSVM formulation

- Given partial ranked lists;
  - create all pairs \((u, v)\)
  - add training data \(\{(x_d, y_d)\}\)

\[
\{(x_u, x_v), 1\} \quad \text{if } u \text{ ranked ahead of } v
\]
\[
\{(x_u, x_v), -1\} \quad \text{otherwise}
\]

- for each type \(t\), solve:

\[
\min_{\Gamma_t} \|\Gamma_t\|_2^2 + \gamma \sum_{d \in D} \epsilon_d
\]

s.t. \(\Gamma'_t(x_d^1 - x_d^2)y_d \geq 1 - \epsilon_d, \forall d \in D\) and \(t_{x_d^1}, t_{x_d^2} = t\)

\[
\epsilon_d \geq 0, \forall d \in D
\]

\[
\Gamma_t(c) \geq 0, \forall c = 1, \ldots, m
\]
Outline

Goal: ranking in directed heterogeneous information networks (HIN) with geo-location

- HINside model
- Parameter estimation
  - via learning-to-rank objectives

Experiments
Experiments I

- Q1: How well does ATR estimation work?
- 2 dataset samples
  - G1: n = 446 physicians of m=3 types, 8537 edges
  - G2: n = 3979 physicians of m=7 types, 93432 edges
- 15 experiments with randomly chosen ATR for G1
- 10 experiments with randomly chosen ATR for G2
- Simulate results based on HINside
  - 1/3 nodes of each type (training), rest as test
Table 1: Mean

<table>
<thead>
<tr>
<th>Method</th>
<th>Type 1</th>
<th>Type 2</th>
<th>Type 3</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>GD-I-NN</td>
<td>0.8852</td>
<td>0.9358</td>
<td>0.9182</td>
<td></td>
</tr>
<tr>
<td>GD-II-NN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RSVM-NC</td>
<td>0.0745</td>
<td>0.0183</td>
<td>0.1818</td>
<td></td>
</tr>
<tr>
<td>GD-I-NC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GD-II-NC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RG</td>
<td>0.0464</td>
<td>0.0852</td>
<td>0.0711</td>
<td></td>
</tr>
<tr>
<td>RO</td>
<td>0.0739</td>
<td>0.0180</td>
<td>0.0643</td>
<td></td>
</tr>
<tr>
<td>INW</td>
<td>0.0852</td>
<td>0.0745</td>
<td>0.0739</td>
<td></td>
</tr>
<tr>
<td>PRANKW</td>
<td>0.0464</td>
<td>0.0183</td>
<td>0.0643</td>
<td></td>
</tr>
</tbody>
</table>

We first investigate the relation of the HIN matrices with small differences inbetween (i.e., swaps in the constraints often estimated a non-negative)

The baselines are unable to capture the ranking by the HIN, despite the increased parameter size (49 vs. 9), where

Interestingly, the ratios of ATR values in this column...
### G2 Test Accuracy - AP@20

<table>
<thead>
<tr>
<th>Method</th>
<th>Type 1</th>
<th>Type 2</th>
<th>Type 3</th>
<th>Type 4</th>
<th>Type 5</th>
<th>Type 6</th>
<th>Type 7</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSVM-NN</td>
<td>0.8367</td>
<td>0.9030</td>
<td>0.9401</td>
<td>0.9639</td>
<td>0.9753</td>
<td>0.9568</td>
<td>0.9362</td>
<td>0.9303</td>
</tr>
<tr>
<td>RSVM-NC</td>
<td><strong>0.8605</strong></td>
<td><strong>0.9361</strong></td>
<td><strong>0.9701</strong></td>
<td>0.9429</td>
<td>0.8829</td>
<td>0.9330</td>
<td><strong>0.9590</strong></td>
<td>0.9263</td>
</tr>
<tr>
<td>GD-I-NN</td>
<td>0.7193</td>
<td>0.8830</td>
<td>0.9074</td>
<td>0.9357</td>
<td>0.8482</td>
<td>0.8812</td>
<td>0.8906</td>
<td>0.8665</td>
</tr>
<tr>
<td>GD-I-NC</td>
<td>0.6999</td>
<td>0.8663</td>
<td>0.9030</td>
<td>0.9015</td>
<td>0.9143</td>
<td>0.8838</td>
<td>0.8710</td>
<td>0.8628</td>
</tr>
<tr>
<td>GD-II-NN</td>
<td>0.8161</td>
<td>0.8978</td>
<td>0.9574</td>
<td>0.9485</td>
<td>0.9441</td>
<td>0.9239</td>
<td>0.9074</td>
<td>0.9136</td>
</tr>
<tr>
<td>GD-II-NC</td>
<td>0.7617</td>
<td>0.8896</td>
<td>0.9465</td>
<td>0.9599</td>
<td>0.9557</td>
<td>0.9177</td>
<td>0.9024</td>
<td>0.9048</td>
</tr>
<tr>
<td>RG</td>
<td>0.5358</td>
<td>0.6483</td>
<td>0.6871</td>
<td>0.6653</td>
<td>0.6796</td>
<td>0.6602</td>
<td>0.6240</td>
<td>0.6429</td>
</tr>
<tr>
<td>RO</td>
<td>0.0029</td>
<td>0.0109</td>
<td>0.0240</td>
<td>0.0494</td>
<td>0.0357</td>
<td>0.0301</td>
<td>0.0326</td>
<td>0.0265</td>
</tr>
<tr>
<td>PRANKW</td>
<td>0.0180</td>
<td>0.0739</td>
<td>0.0464</td>
<td>0.0852</td>
<td>0.0745</td>
<td>0.0183</td>
<td>0.1818</td>
<td>0.0711</td>
</tr>
<tr>
<td>INW</td>
<td>0.2143</td>
<td>0.2808</td>
<td>0.3053</td>
<td>0.1326</td>
<td>0.2725</td>
<td>0.3946</td>
<td>0.2555</td>
<td>0.2651</td>
</tr>
</tbody>
</table>

- A: RankSVM with non-negative (-NN) ATR constraints works well
Q2: How well does HINside reflect real world?

Dataset: author graph of collaborations from 4 areas publicly available at http://web.engr.illinois.edu/~mingji1/DBLP_four_area.zip

Crawled institution (location) for ~11K authors

- Locations from 72 unique countries, 6 continents

No agreed-upon ranking of researchers (even within the same area)

Compare/contrast HINside, Pagerank, h-index

- Pagerank: no location, just co-authorship
- h-index: not co-authorship but citations
HIN\textsuperscript{side}, Pagerank, h-index

Example cases for which model differ significantly:

<table>
<thead>
<tr>
<th>Name</th>
<th>Area</th>
<th>Institution</th>
<th>h</th>
<th>P</th>
<th>HIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moshe Vardi</td>
<td>DB</td>
<td>Rice U.</td>
<td>87</td>
<td>165</td>
<td>17</td>
</tr>
<tr>
<td>Michael R. Lyu</td>
<td>IR</td>
<td>CUHK</td>
<td>67</td>
<td>83</td>
<td>1</td>
</tr>
<tr>
<td>Andreas Krause</td>
<td>ML</td>
<td>ETH Zurich</td>
<td>45</td>
<td>291</td>
<td>4</td>
</tr>
</tbody>
</table>

Carnegie Mellon
Summary

Goal: ranking nodes in directed heterogeneous information networks (HIN) with geo-location

- Designed **HINside model**, incorporating
  - (1) relation **strength**, (2) pairwise **distance**, (3) neighbors’ authority scores, (4) authority transfer rates (**ATR**) between different types of nodes, and (5) **competition** due to co-location
  - Location info dictates (2) and (5)
  - **Closed form formula**

- Derived **parameter (ATR) estimation algorithms**
  - HINside lends itself to learning the ATR via **learning-to-rank objectives**
  - Proposed and studied two: (i) RankSVM based, and (2) pairwise rank-ordered log likelihood
Thanks!

Paper, Code, Data, Contact info:
www.cs.cmu.edu/~lakoglu
https://github.com/abhimm/HINSIDE