Ranking in Heterogeneous Networks with Geo-Location Information

Abhinav Mishra
Amazon

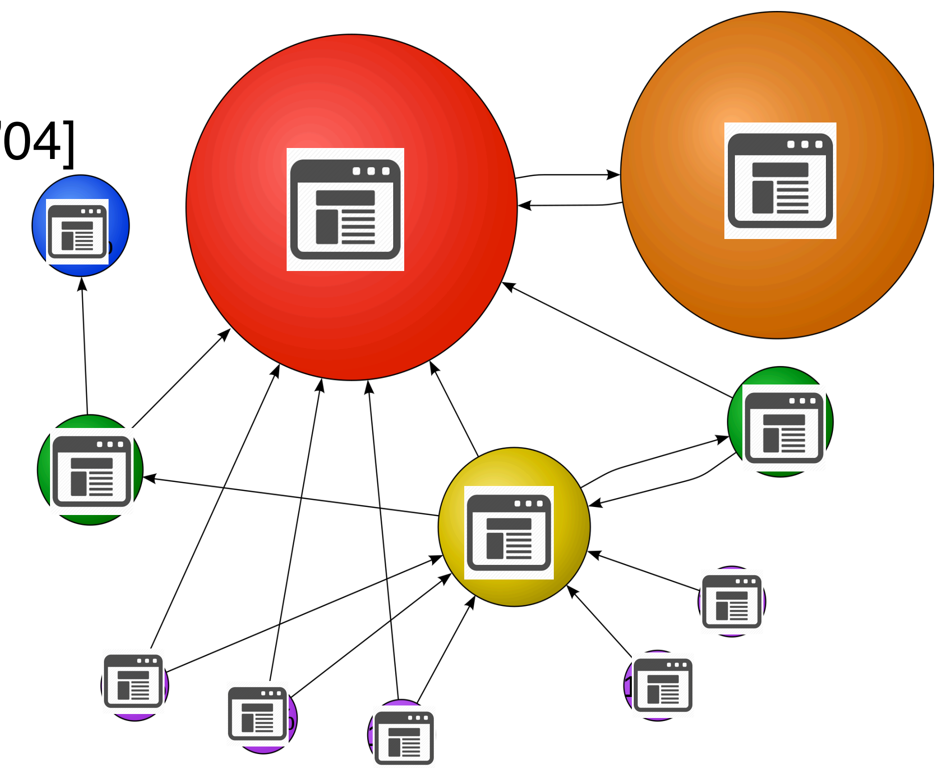
Leman Akoglu
CMU

Carnegie Mellon University Heinz College

SIAM SDM 2017
Houston, Texas

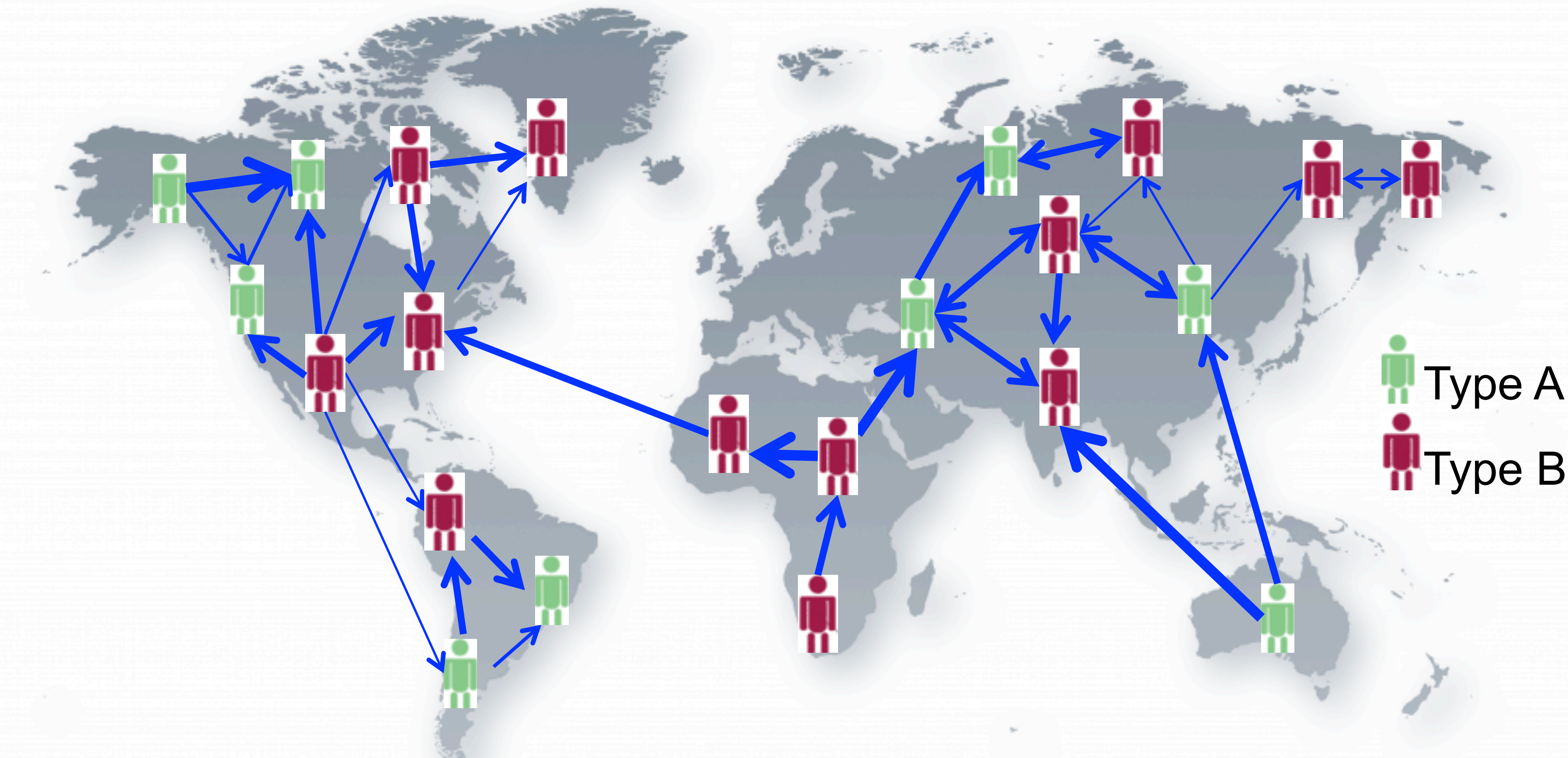
Ranking in networks

- 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]
  - …



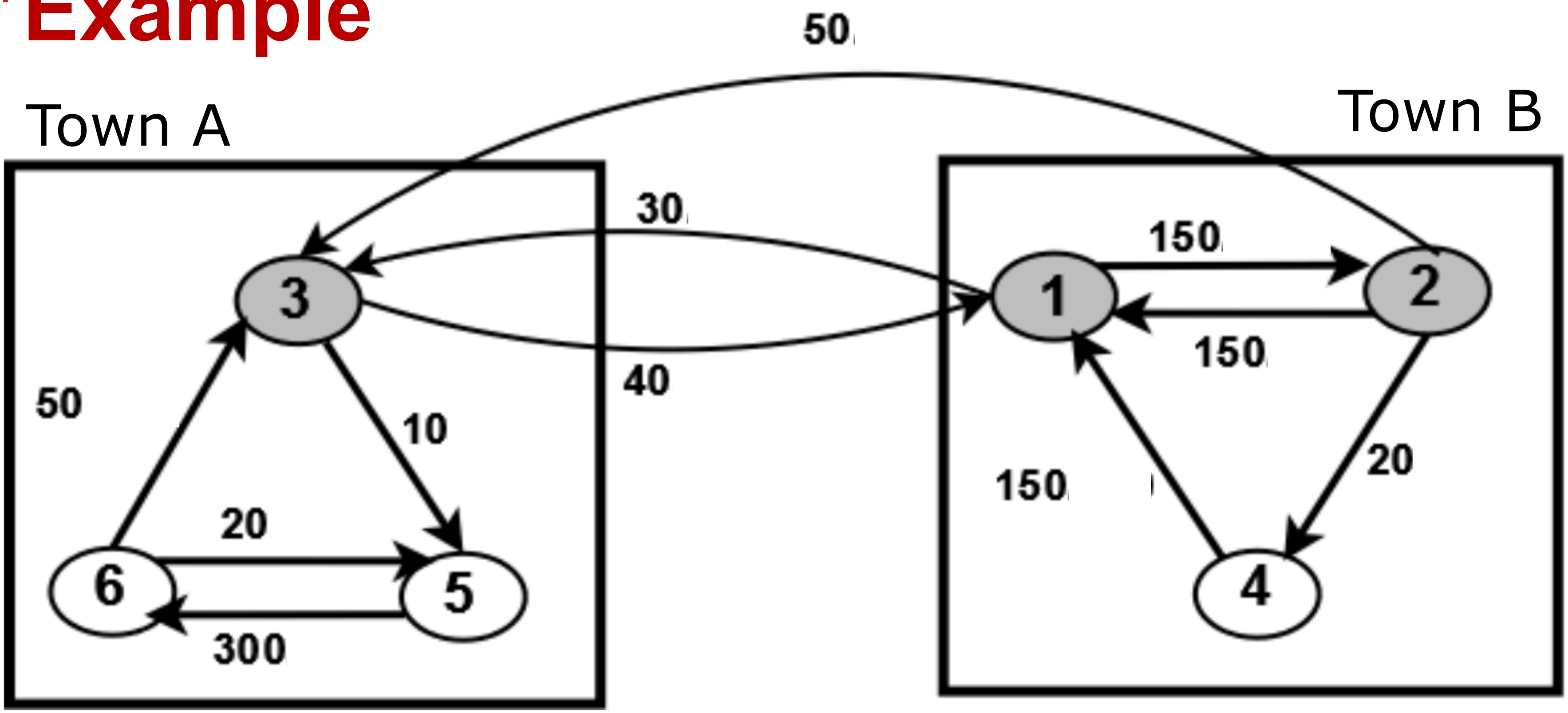
Ranking in rich networks

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



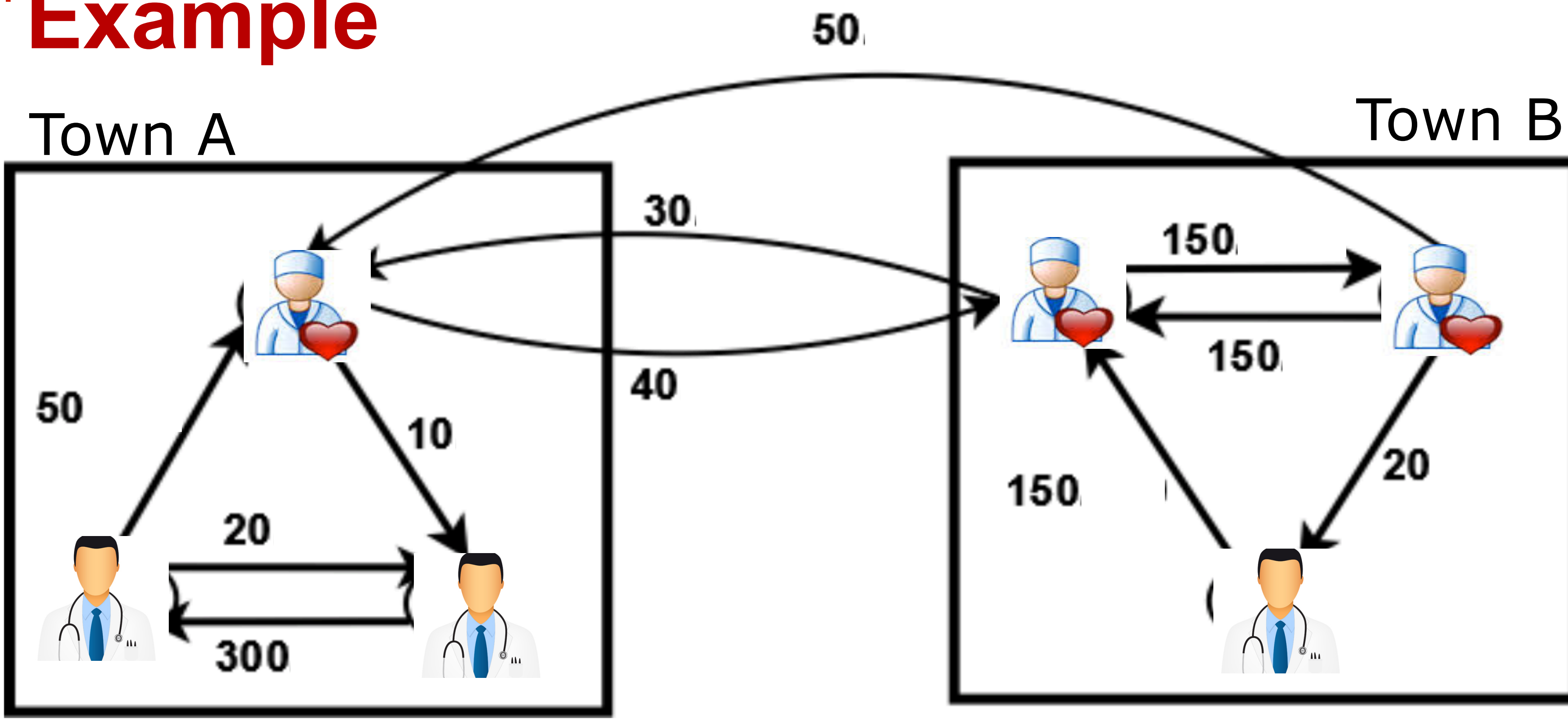
- Different types of nodes ranked separately

Example



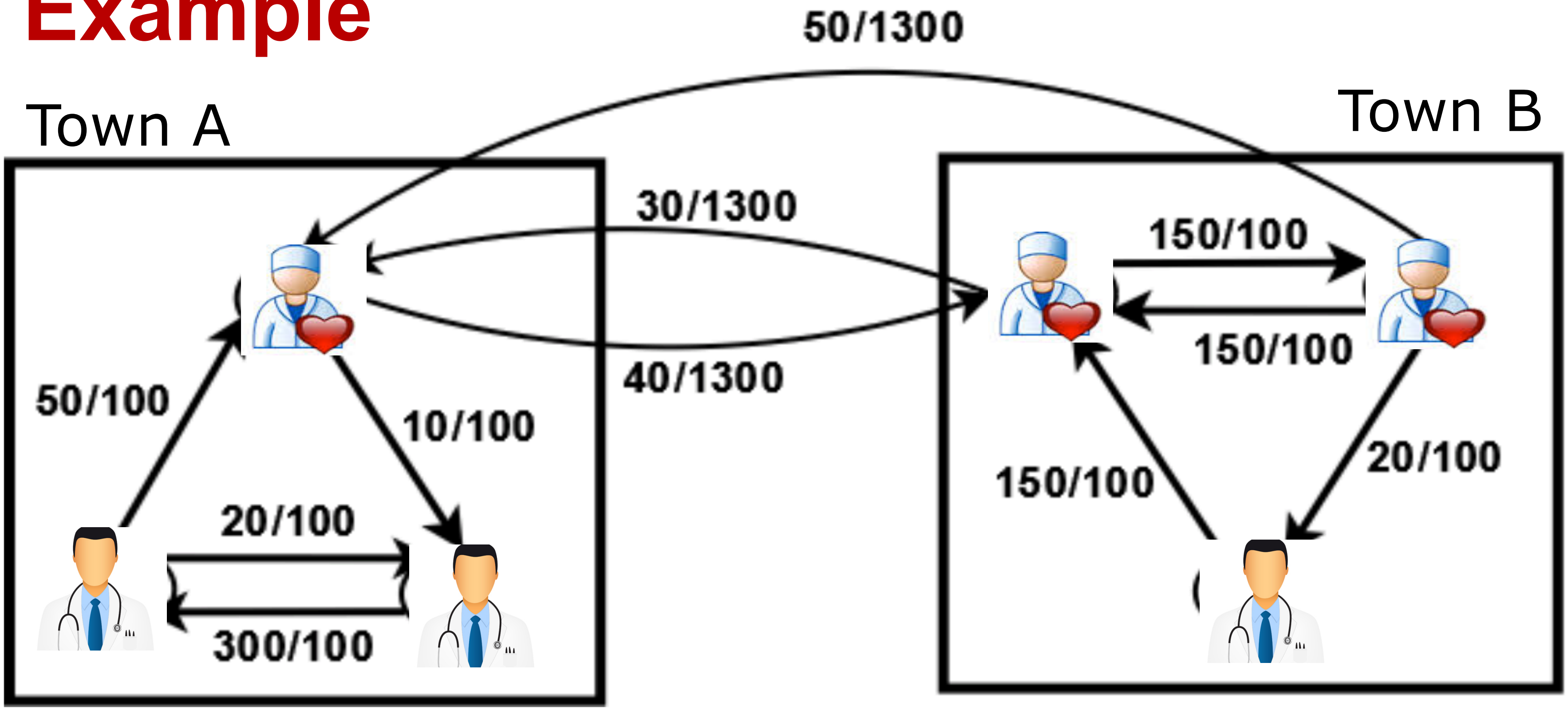
Weighted medical referral network (directed)

Example



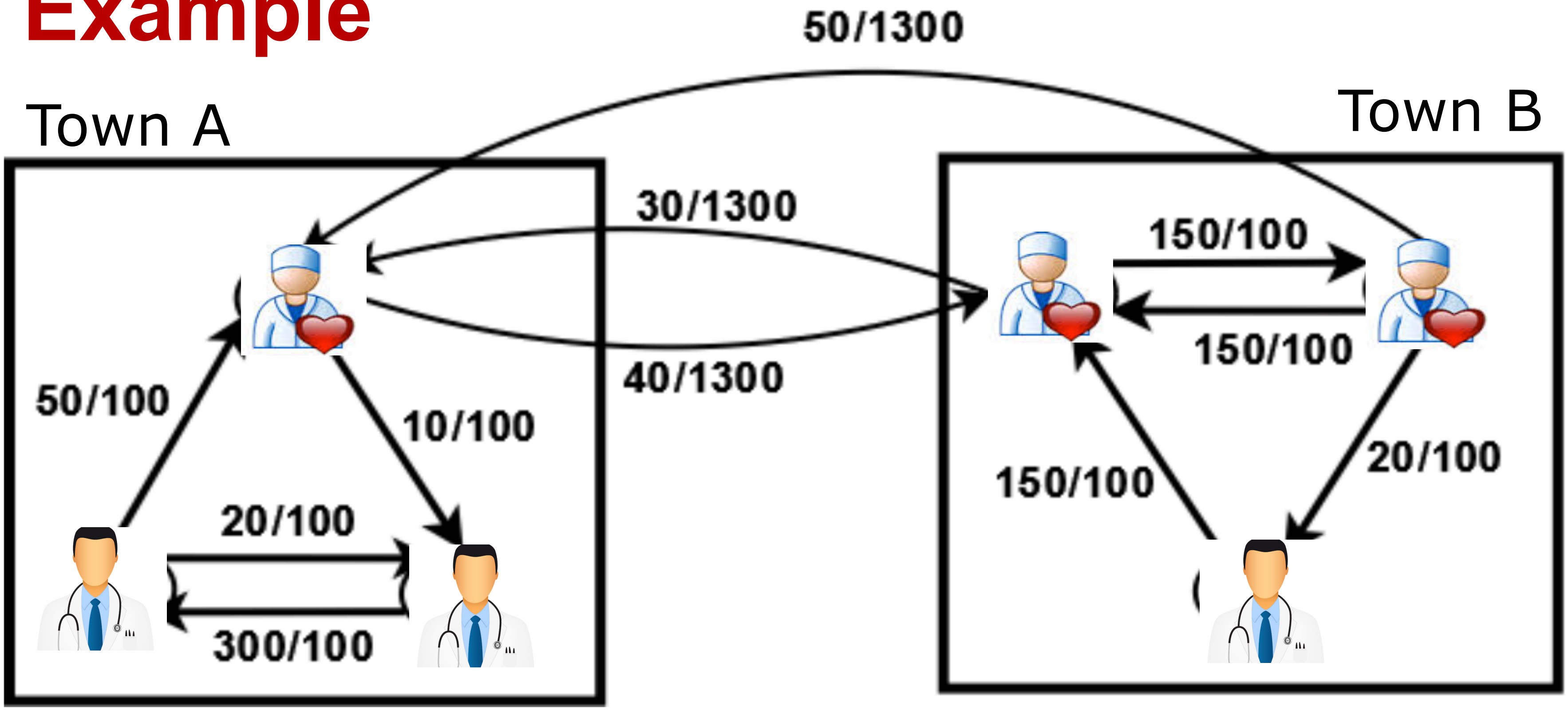
Weighted medical referral network (directed)
+ physician expertise

Example



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

Example

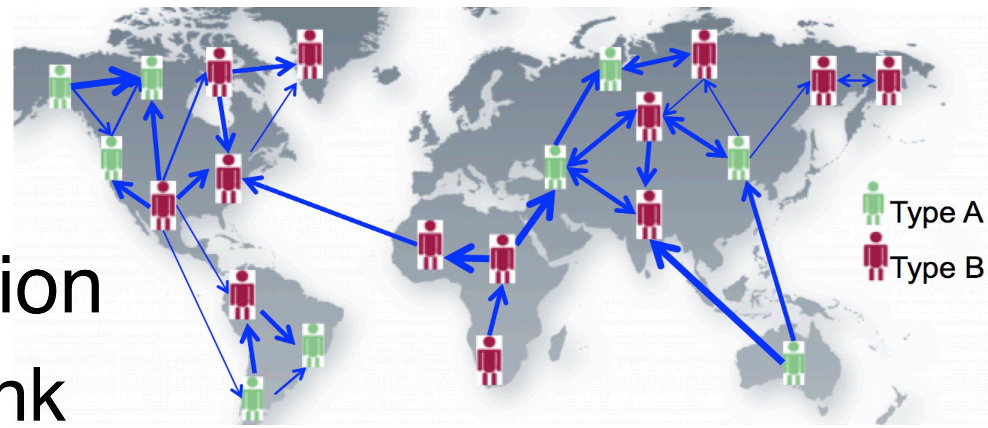


Ranking Problem: Which are the top k nodes of a certain type?

e.g.: Who are the best cardiologists in the network, in my town, etc.?

Outline

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

$$W(i,j) = \log(w(i,j) + 1)$$

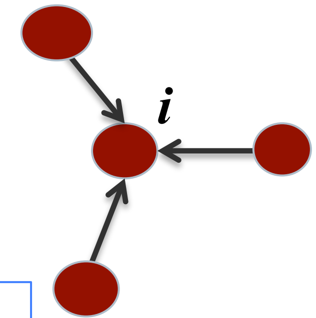
  - pair-wise distances

$$D(i,j) = \log(d(l_i, l_j) + 1)$$

$$M = W \odot D \tag{3.1}$$

HINside model

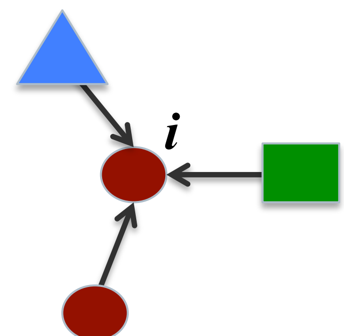
- In-neighbor authority



$$r_i = \sum_{j \in \mathcal{V}} M(j,i)\, r_j \qquad (3.2)$$

r_i : authority score of node i

- Authority Transfer Rates (ATR)

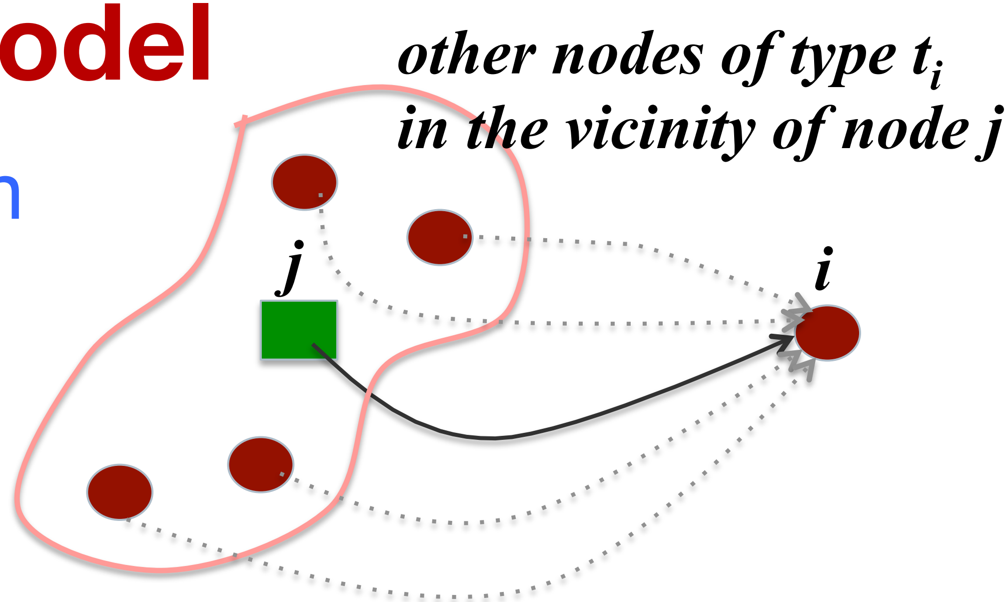


$$r_i = \sum_{j \in \mathcal{V}} \Gamma(t_j, t_i)\; M(j,i)\, r_j \qquad (3.3)$$

t_i : type of node i

HINside model

- Competition



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

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

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

Closed-form solution

- Authority scores vector **r** written in closed form as (& computed by power iterations)

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

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

$$E = TT'$$

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 m² authority transfer rates (ATR)

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

  - r_i as a vector-vector product

$$r_i = \sum_t \Gamma(t, t_i) \sum_{j:t_j = t} [M(j,i)(r_j + \sum_{v:t_v = t_i} N(v,j)\, r_v)]$$

$$r_i = \sum_t \Gamma(t, t_i) X(t, i)$$

$$= \Gamma'(t_i, :) \cdot X(:, i) = \mathbf{\Gamma}'_{t_i} \cdot \mathbf{x}_i$$

$$= f(\mathbf{x}_i) = < \mathbf{w}, \mathbf{x}_i >$$

An alternating optimization scheme:

- $\Gamma \longrightarrow \mathbf{r} \longrightarrow X \xrightarrow{\text{estimate}} \Gamma$

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

- Randomly initialize Γ^0 , $k = 0$
- Compute authority scores **r** using Γ^0
- **Repeat**
  - X^k ← compute feature vectors using **r**
  - Γ^{k+1} ← learn new parameters by learning-to-rank
  - compute authority scores **r** using Γ^{k+1}
- **Until** convergence

An alternating optimization scheme:

- $\Gamma \longrightarrow \mathbf{r} \longrightarrow X \xrightarrow{\text{estimate}} \Gamma$

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

- Randomly initialize Γ^0 , $k = 0$
- Compute authority scores **r** using Γ^0
- **Repeat**
  - X^k ← compute feature vectors using **r**
  - Γ^{k+1} ← learn new parameters by learning-to-rank
  - compute authority scores **r** using Γ^{k+1}
- **Until** convergence

RankSVM formulation

Cross-entropy based objective by gradient descent

- Given partial ranked lists;
  - create all pairs (u, v)
  - add training data $\{((\mathbf{x}_d^1, \mathbf{x}_d^2), y_d)\}_{d=1}^{|\mathcal{D}|}$

    $((\mathbf{x}_u, \mathbf{x}_v), 1)$ if u ranked ahead of v

    $((\mathbf{x}_u, \mathbf{x}_v), -1)$ otherwise
  - for each type t, solve:

$$
\begin{aligned}
&\min_{\boldsymbol{\Gamma}_t} \ ||\boldsymbol{\Gamma}_t||_2^2 + \gamma \sum_{d \in \mathcal{D}} \epsilon_d \\
&\text{s.t. } \boldsymbol{\Gamma}_t'(\mathbf{x}_d^1 - \mathbf{x}_d^2) y_d \geq 1 - \epsilon_d, \ \forall d \in \mathcal{D} \text{ and } t_{\mathbf{x}_d^1}, t_{\mathbf{x}_d^2} = t \\
&\qquad \epsilon_d \geq 0, \ \forall d \in \mathcal{D} \\
&\qquad \boldsymbol{\Gamma}_t(c) \geq 0, \ \forall c = 1, \ldots, m
\end{aligned}
$$

Outline

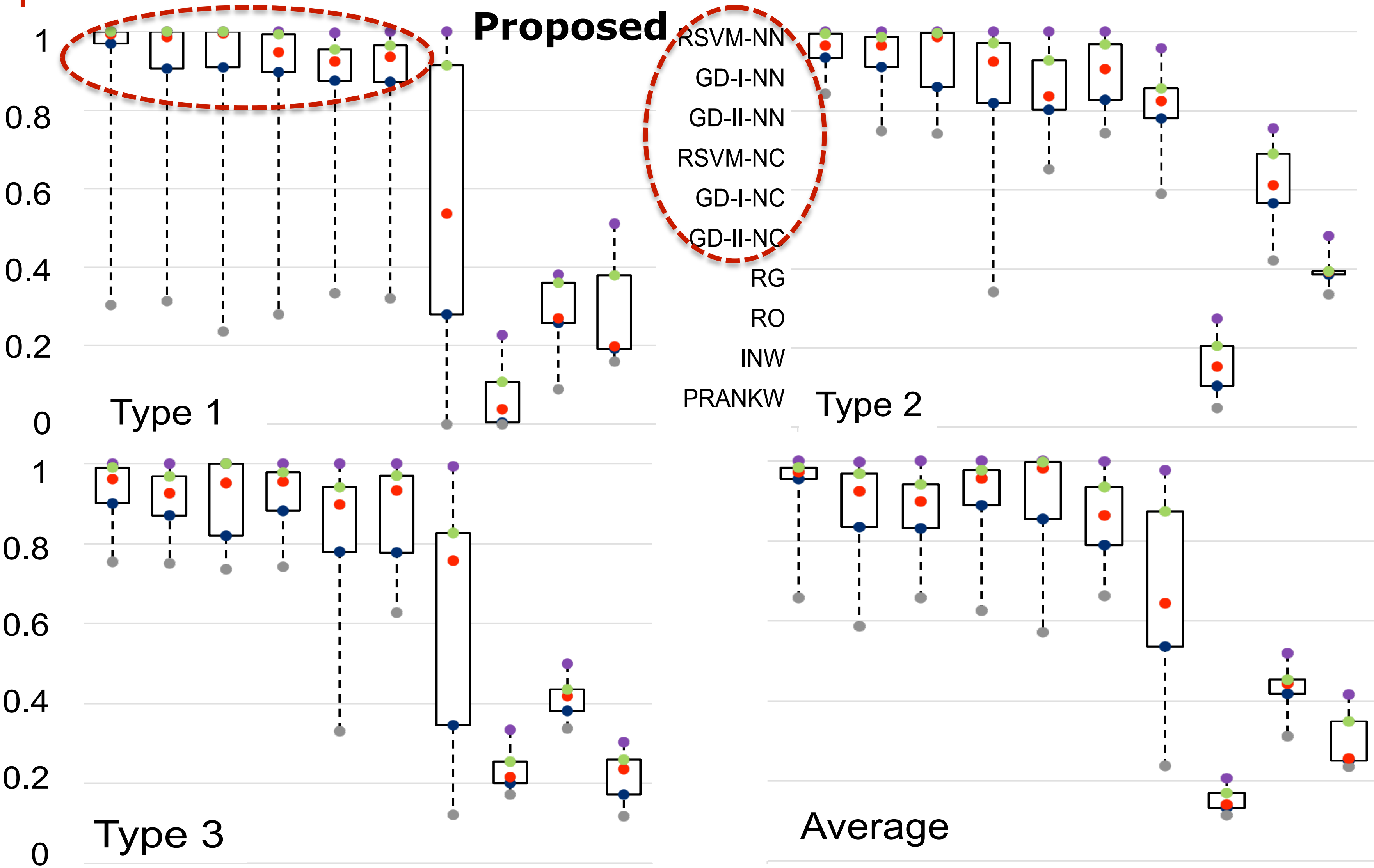
Goal: ranking in directed heterogeneous information networks (HIN) with geo-location

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

Experiments I

- Q1: How well does ATR estimation work?
- Datasets: physician referral data for years 2009–2015 publicly available at https://questions.cms.gov/faq.php?faqId=7977
- 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

G1 Test Accuracy - AP@20



G2 Test Accuracy - AP@20

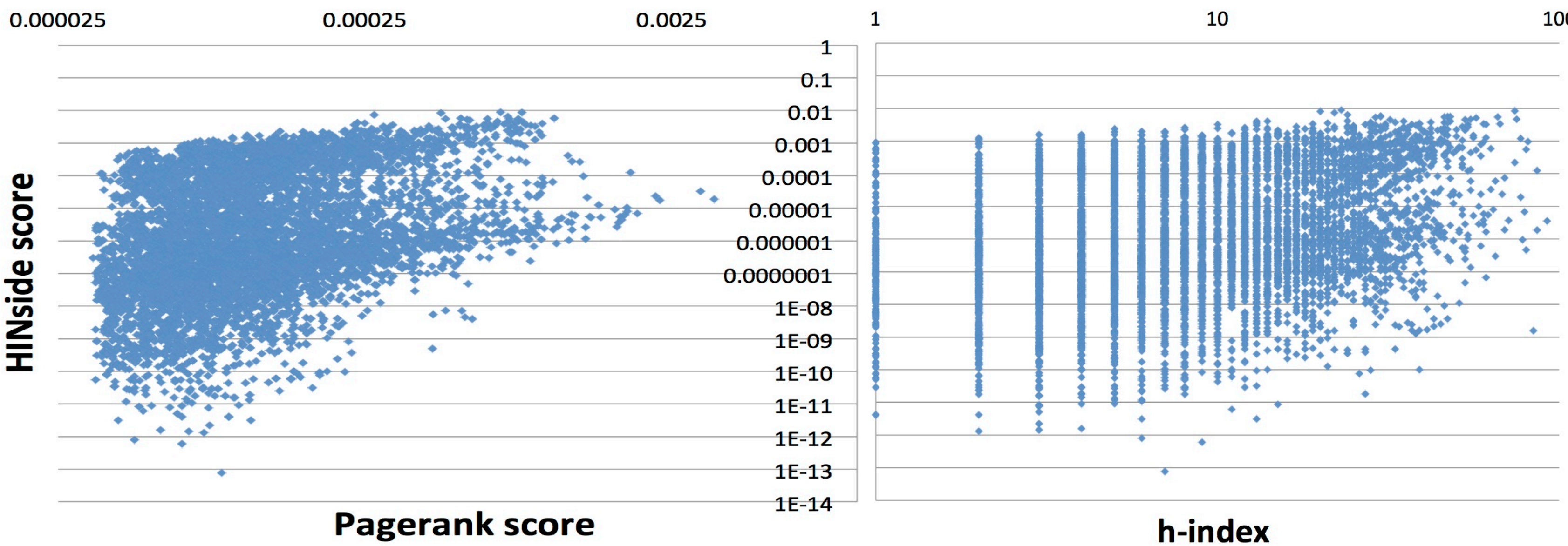
Method	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7	Average
RSVM-NN	0.8367	0.9030	0.9401	**0.9639**	**0.9753**	**0.9568**	0.9362	**0.9303**
RSVM-NC	**0.8605**	**0.9361**	**0.9701**	0.9429	0.8829	0.9330	**0.9590**	0.9263
GD-I-NN	0.7193	0.8830	0.9074	0.9357	0.8482	0.8812	0.8906	0.8665
GD-I-NC	0.6999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

- A: RankSVM with non-negative (-NN) ATR constraints works well

Experiments II

- Q2: How well does HINside reflect real world?
- Dataset: author graph of collaborations from m=4 areas publicly available at http://web.engr.illinois.edu/~mingji1/DBLP_four_area.zip
- Crawled institution (location) for n= ~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

HINside, Pagerank, h-index



Example cases for which model differ significantly:

Name	Area	Institution	h	P	HIN
Moshe Vardi	DB	Rice U.	87	165	17
Michael R. Lyu	IR	CUHK	67	83	1
Andreas Krause	ML	ETH Zurich	45	291	4

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