10/30/2020

① Missing Node Labels.

② Missing Edges.

Why do we have missing data?

- In Medicine, diff collection standards.
- If "Optional" field on a survey... (some nodes opt to not disclose label).
- Classified Data - unreleased (privacy issues)
- Sampling → got network, but not all the labels.
  - Partial visibility
- Data Cleaning, Wrangling issues.

Heuristic to fill in the gaps.

↓

M.L. models that we call "semi supervised"

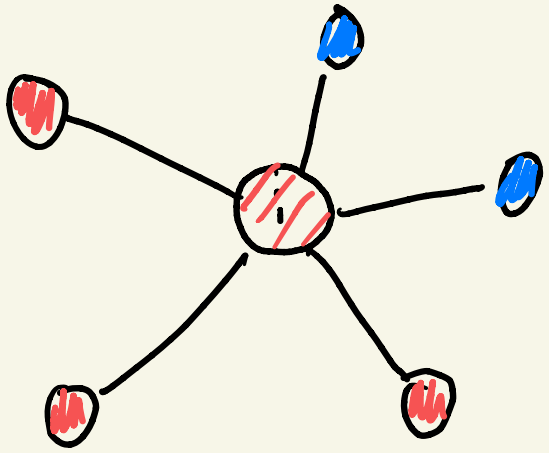
Goal: impute / generalize known labels → unknown labels.

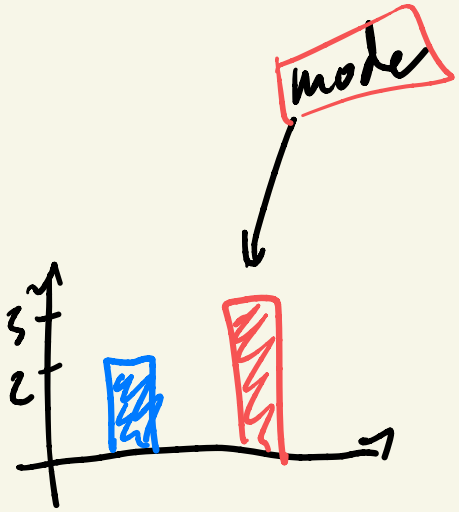
"Guilt By Association"

GBA:

- For a node i with no label, "guess" that its missing label is the mode of its neighbors' (known) labels.
- Break ties U.A.R.

Local Neighborhood.





- When is this likely to work well? [Network Gerrymandering??]
- Could be tough if there are too many unlabeled nodes.
- Assumes Homophily.

Test Bed for How this heuristic performs:

1. take a network w/ full labels.
2. chuck out U.A.R. fraction $1-f$ of labels (f remain)
3. Apply heuristic
4. Evaluate my predictions
5. Repeat, redrawing the fraction.

Monophily:

Tendency to link with others of only one type.

Altenberger + Ugander.

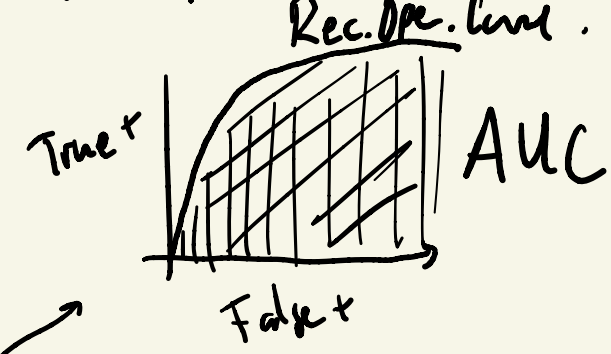
Link Prediction

Test Bed:

1. Reveal only f of links
2. Apply alg'm or heuristic:
   - Ranked list, starting from most likely non-edge that we think should be an edge, and decreasing.
3. Evaluate by moving down the list... correct vs incorrect as a function of how far down list...

magic (3 examples, next slide)

- score every (i,j) in network where $A_{ij} = 0$ (non-edge)
- sort!



3 methods for Link Prediction.

score(i,j)

① ChungLu / C.M.

Degree-Product.

$score(i,j) = k_i k_j$

② Normalized Common Nbrs.

Let N_i be neighbors of i

$$score(i,j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|}$$

$\cap$: cap

$\cup$: cup

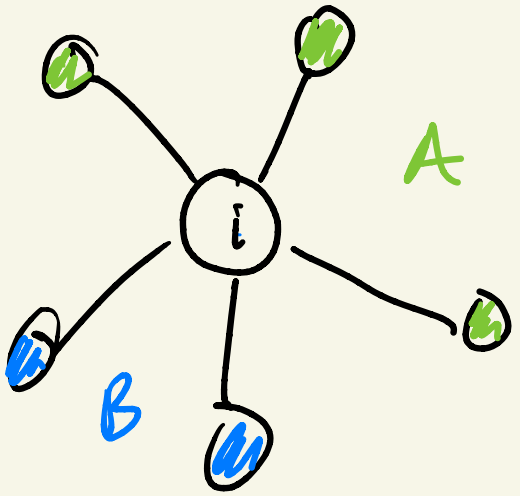
Jaccard Coeff. → Botany. Triadic Closure

③ Shortest Path.

$$score(i,j) = \frac{1}{\sigma(i,j)}$$

$\sigma(i,j)$: length of shortest path between i and j.

- E.R. Network. $G(n,p)$
- 2 labels.
- Fraction class I = q
- f fraction labels



$\Pr(i \text{ gets label } A)$

$= \Pr(k_A > k_B) + \frac{1}{2}\Pr(k_A = k_B) = \;\underline{\qquad}\;\underline{\qquad}\; \overset{?}{\geq} q$

sum

Degree $k_i = \text{Bin}(n,p)$

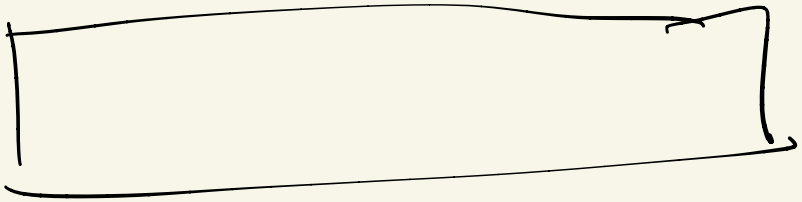
$k_A = \text{Bin}(k_i, q)$

$k_B = \text{Bin}(k_i, 1-q)$

$\Pr(k_A > k_B) = \Pr(k_A - k_B > 0)$ ← sum.

$$Pr(k_A - k_B > 0) = \sum_t Pr(k_A - t > 0)\, Pr(k_B = t)$$

$$\sum_t Pr(k_A - k_B > 0 \mid k_B = t)\, Pr(k_B = t)$$

$$Pr(k_A = k_B) = \sum_s$$  ?