Web Structure Mining

Graph-based Learning: Principles and Applications

Li Xiaoli

NUS National University of Singapore
ISS INSTITUTE OF SYSTEMS SCIENCE

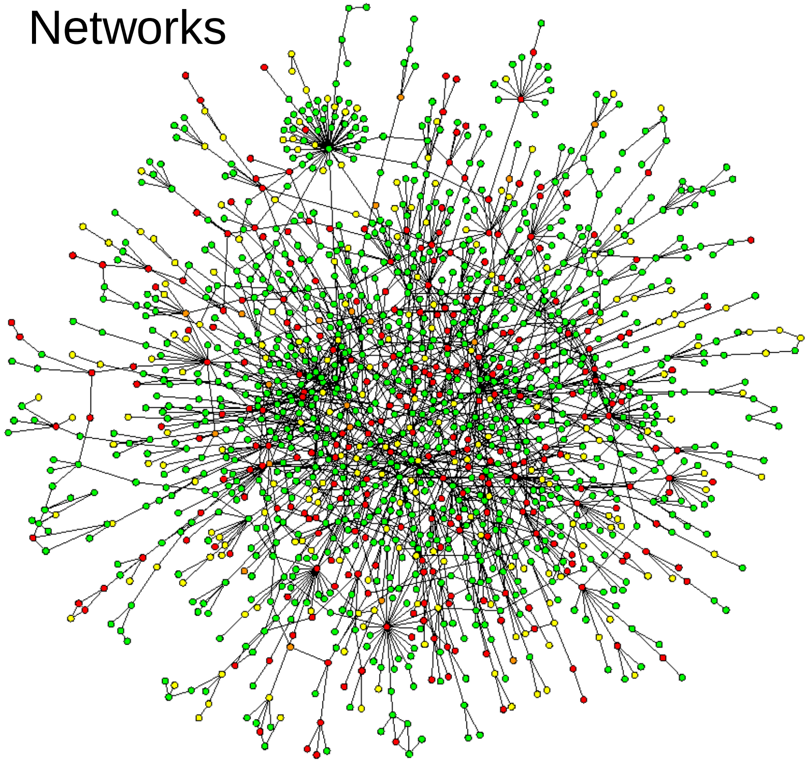
Outlines

PART I: Insights and Principle

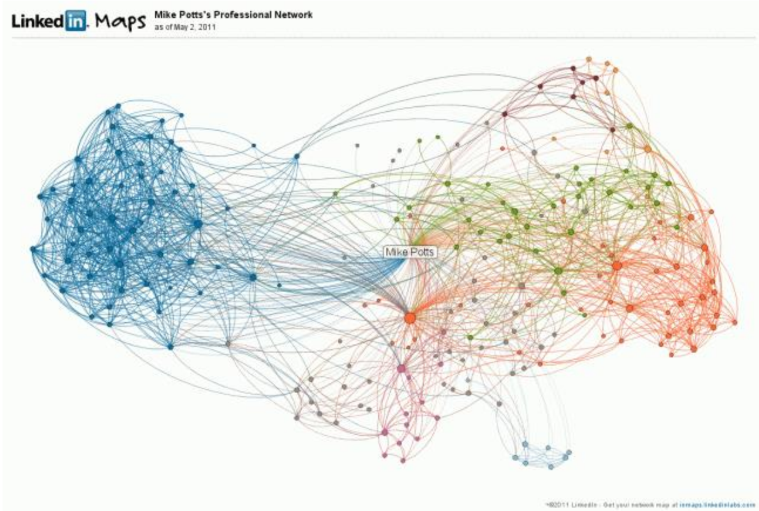
PART II: Applications [Briefly]

Networks are Ubiquitous

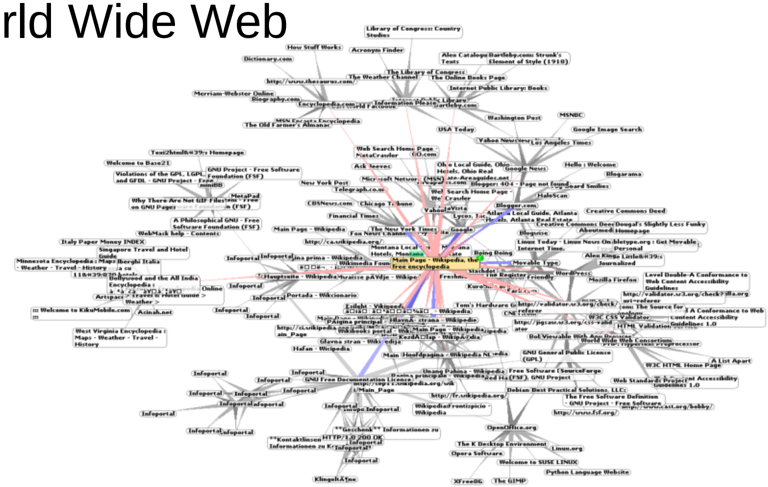
Protein-Protein Interaction Networks



Social Networks

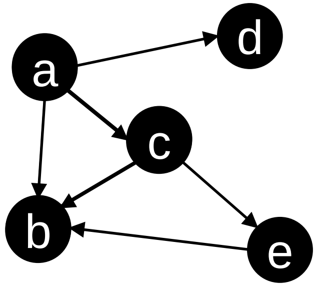


World Wide Web



Graph Representation

- ***Graph*** is the data structure to model networks
- $G = (V, E)$
  - V: the set of nodes
  - E: the set of edges (links), directed or undirected
- Example
  - $V = \{a, b, c, d, e\}$
  - $E = \{\langle a, b\rangle, \langle a, c\rangle, \langle a, d\rangle, \langle c, b\rangle, \langle c, e\rangle, \langle e, b\rangle\}$



Graph-based Learning: should be applied to both directed and undirected graphs

Part I: Insights and Principle

Learning on a graph

- Classification on an affinity graph [Zhu et al, ICML03]
- Input: Affinity graph
  - Two nodes are connected if they are *similar*
  - W_{ij}: an element in a matrix W to indicate the strength of similarity between i, j

Weighted graph. We can perform K-NN sparsification
W can be either engineered (cosine similarity) or
natural (links in social network)

- Task: Classification
  - Given class labels of some nodes
  - Predict those unknown labels
  - Example:
    - Which node has a label?
    - Which node is unknown, unlabeled, or does not have a label?

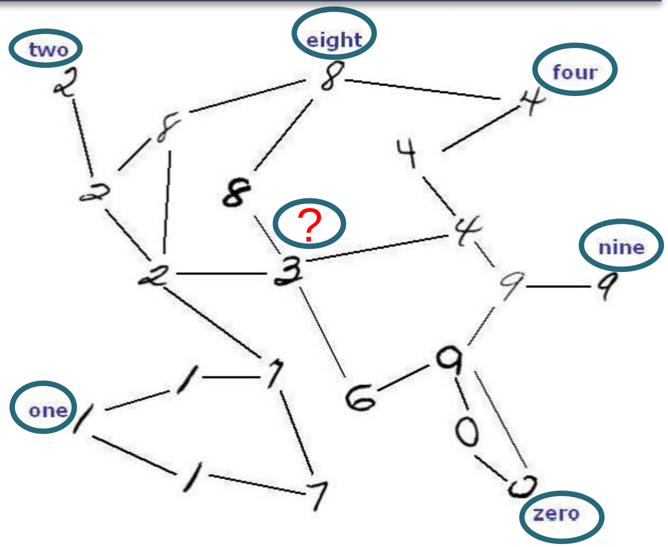
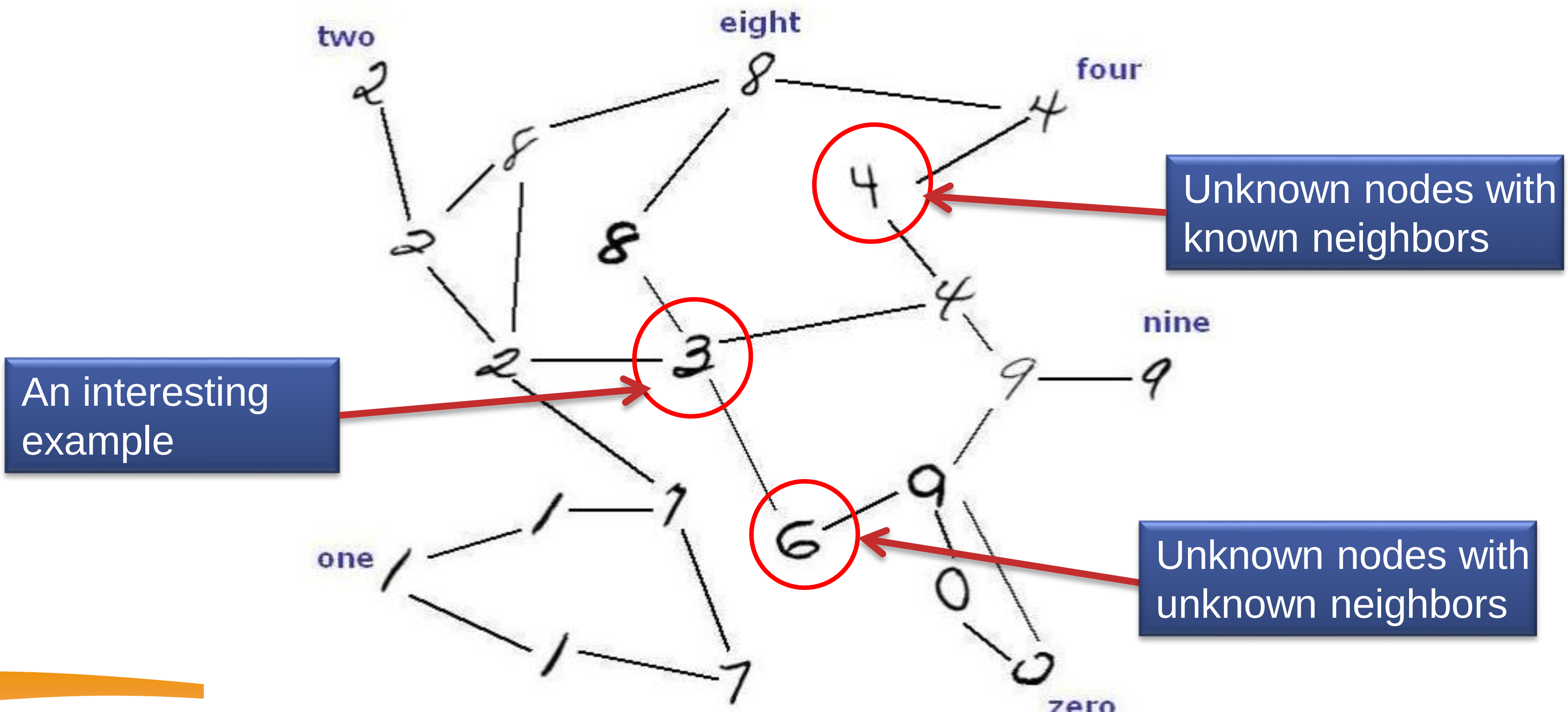


Image Credit: Xiaojin Zhu, etal, ICML'03

Key Intuition

- A link between two nodes shows they are similar
- Similar nodes should have similar class labels

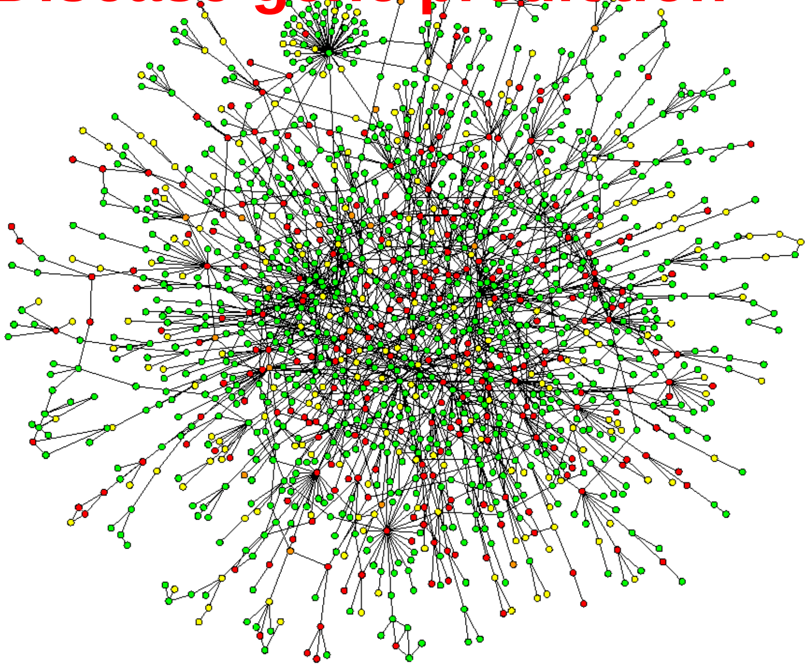


Network Node Classification

Protein-Protein Interaction Networks

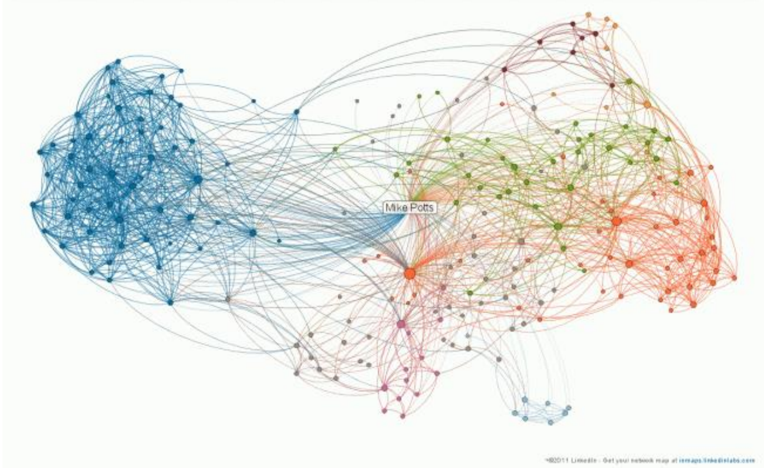
Protein function prediction

Disease gene prediction



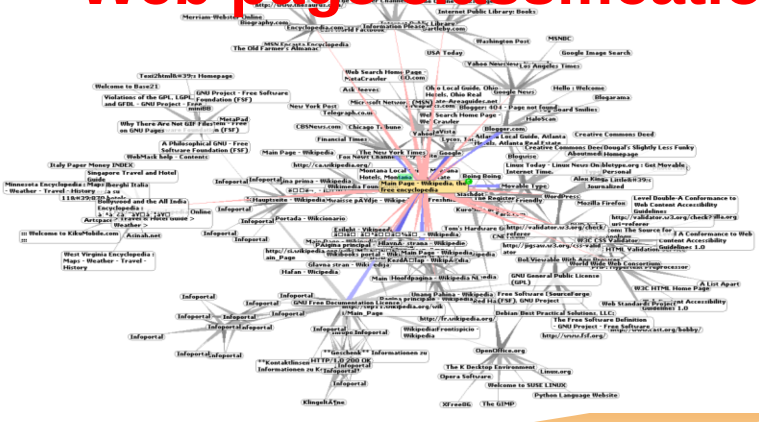
Social Networks

Customer classification



World Wide Web

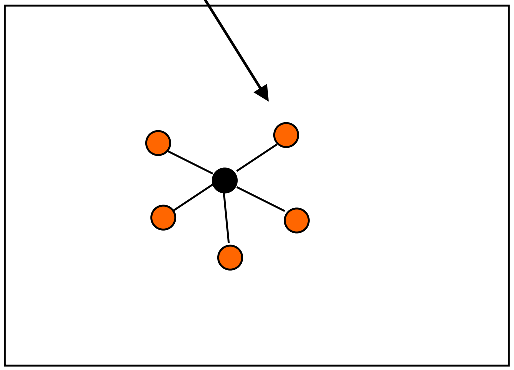
Web page classification



Protein Function Prediction - Real-world Example in Biological networks

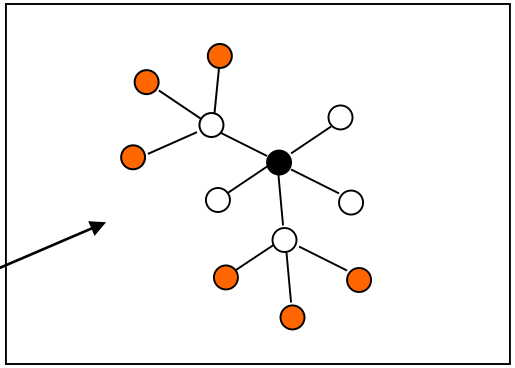
- Direct functional association:
  - Interaction partners of a protein are likely to share functions with it
- Indirect functional association
  - Proteins that share interaction partners with a protein may also likely to share functions with it

Level-1 neighbor



- 59.2% proteins in PPI network share some function with level-1 neighbors
- 70.7% proteins in PPI network share some function with level-1 or level 2 neighbors

Level-2 neighbor



NUS National University of Singapore | ISS INSTITUTE OF SYSTEMS SCIENCE

A key research publication

- Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. **Semi-supervised learning using Gaussian fields and harmonic functions**. In *The 20th International Conference on Machine Learning (ICML)*, 2003. **ICML 10-Year Classic Paper Prize.**
- A graph-based semi-supervised learning algorithm that creates a graph over labeled and unlabeled examples. More similar examples are connected by edges with higher weights. The intuition is for the labels to propagate on the graph to unlabeled data. The solution can be found with simple matrix operations, and has strong connections to spectral graph theory. [ps.gz] [pdf] [Matlab code] [data]

Formalization

- Assuming binary classes $\{0,1\}$
- $f_i \in [0,1]$: prediction on a node i
- Minimizing the following:

Class labels. e.g.
0: non-customer,
1: customer

Share same label

4 4

$$E(f) = \frac{1}{2}\sum_{ij} W_{ij}(f_i - f_j)^2$$

f_i: confidence score of node *i to be label 1*.
If a node *i* is a labelled node, then f_i belongs to {0, 1}.
If a node *i* is unlabelled/unknown: f_i needs to be inferred

Input W_{ij}: weight/similarity between node *i* and *j*
E(f) means if node i is similar to node j, then their confidence score is similar (energy function of Gaussian random field: See reference paper)

Minimizing E(f)

- How to minimize E(f)?
- There are *iterative updating method* and *matrix based method*. However, iterative updating method is much more *efficient* than matrix method, as matrix method needs more expensive operations, e.g. matrix inverse
- Then how to do iterative method?

Minimizing E(f)

- Iterative updating
  - Confidence score Initialization (0 iteration)

$$f_i^{(0)} = \begin{cases} 1 & \text{node } i \text{ is labeled as class 1} \\ 0 & \text{else} \end{cases}$$

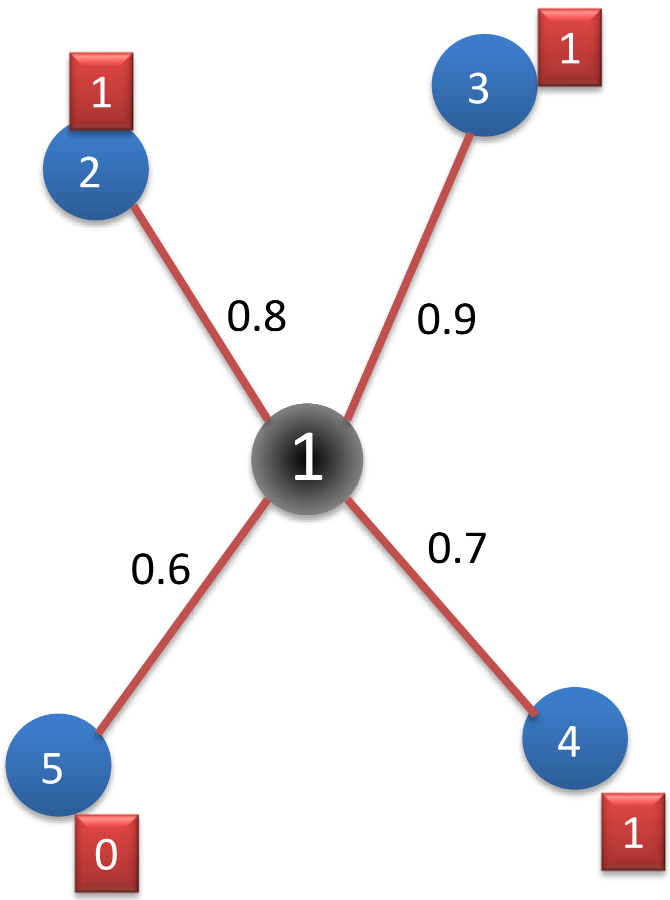
  - For following iteration, i.e. $k = 1,2,3,\ldots$

$$f_i^{(k)} = \begin{cases} 1 & \text{node } i \text{ is labeled as class 1} \\ 0 & \text{node } i \text{ is labeled as class 0} \\ \dfrac{\sum_j W_{ij} f_j^{(k-1)}}{\sum_j W_{ij}} & \text{else} \end{cases}$$

Set $\nabla E(f) = 0$

- Every node's confidence score is kind of the weighted average confidence score of its neighbors
- We ignore those neighbors without labels {0,1} or soft-labels [0,1]

Example of Iterative Updating



- 0 Iteration $f_2^{(0)}=1,\quad f_3^{(0)}=1,$ $f_4^{(0)}=1, f_5^{(0)}=0$

  weights $W_{12}=0.8,\ W_{13}=0.9,$ $W_{14}=0.7,\ W_{15}=0.6$

- Then 1 iteration for node 1
- $f_1^{(1)} = \frac{0.8*1+0.9*1+0.7*1+0.6*0}{0.8+0.9+0.7+0.6}$

  =2.4/3

  =0.8

Starting from node 1, randomly move to neighbors (take weights into consideration, i.e. high weight gets higher prob), there is 0.8 probability to reach a node with label 1

Matrix method

- *Input: W*
- $D = diag(d_i)$ where $d_i = \sum_j w_{ij}$ (row sum, and add into diagonal of D)
- $P = D^{-1}W$ [divided W and P into 4 blocks]
- $W = \begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix}$ $f = \begin{bmatrix} f_l \\ f_u \end{bmatrix}$ $P = \begin{bmatrix} P_{ll} & P_{lu} \\ P_{ul} & P_{uu} \end{bmatrix}$

$$f_u = (D_{uu} - W_{uu})^{-1} W_{ul} f_l = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{uu}})^{-\mathbf{1}} \boldsymbol{P}_{\boldsymbol{ul}} \boldsymbol{f}_{\boldsymbol{l}}$$

Iterative updating – further remarks

- In iterative process, the confidence score for any node *i* in *k-th* iteration $f_i^{(k)}$ could be a value within [0,1], instead of {0,1}. We basically use the soft labels (kind of uncertain) of *i's* neighbors to update *i's* score
- Given limited labelled data (nodes with known labels, i.e. 0 or 1), in the beginning, it is highly likely that some (or even all) neighbors of a node *i* have no existing scores. However, the influence from labelled data will eventually propagate to these unknown nodes (indirectly, step by step).
- The algorithm will converge eventually – confidence scores for all the nodes do not change or have very little changes between two iterations.

NUS National University of Singapore | ISS INSTITUTE OF SYSTEMS SCIENCE

Significance

- An instance of ***semi-supervised*** learning
- Our prediction model is based on both labeled and unlabeled data.
  - We use labelled data to predict the soft labels (confidence scores) of unlabeled/unknown nodes
  - Labels will propagate across the whole network (some nodes need a couple of times to get its temporary scores).

Labels + Structure in data = Semi-Supervised

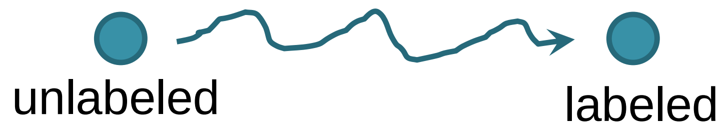
Affinity Graph

NUS National University of Singapore | ISS INSTITUTE OF SYSTEMS SCIENCE

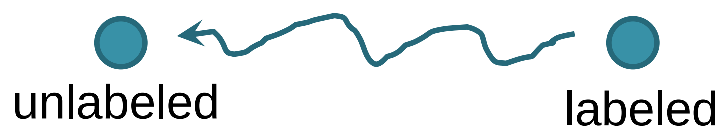
Random Walk Interpretations

- Backward style
  - Starting from an unlabelled node
  - Moves to a random neighbour until hitting a labelled node
  - Will try the same process multiple times and compute

$$f(i) = P(\text{hit a node labeled 1} | \text{starting from } i)$$



- Forward style
  - Starting from a labelled node
  - Moves to a neighboring node randomly
  - Probability of reaching node i (rough idea of PageRank)



- More in [Agarwal et al WWW10], [WSDM11], [ICML14]

Part II: Applications

Applications of graph-based learning

- Spatiotemporal entity linking [TACL14]
- Pattern-based relation extraction [WSDM11]
- Other applications
  - Query classification on query graph [SIGIR12]
  - Semantic ranking on an entity graph [ICDE13]
  - Many more…

Application 1:

Spatiotemporal Entity Linking on Microblogs

Y. Fang and M.-W. Chang. Entity Linking on Microblogs with Spatial and Temporal Signals. In *TACL* 2(Oct), 2014, pp. 259−272.

What is it?

Entity Linking in Microblogs: Map entity mentions in a message (e.g. a tweet) into predefined entities (e.g. entries in Wikipedia) or other knowledge base.

US secretary of state Clinton is hospitalized due to …

http://en.wikipedia.org/wiki/Hillary_Rodham_Clinton

http://en.wikipedia.org/wiki/United_States

Why is it important?

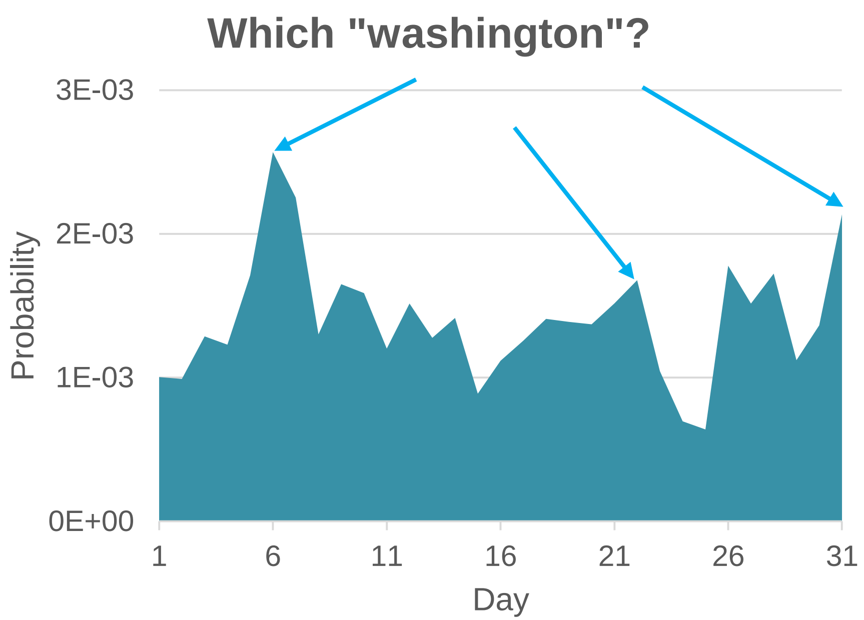
- Motivation: intelligence gathering
- Word-based matching is ineffective due to **ambiguity**

"Washington"?

"Spurs"?

Why is it important?

- Motivation: intelligence gathering
- Word-based matching is ineffective due to **ambiguity**



1) Different peaks → Different entities?
2) A single peak → A mixture of entities?

Observation & Intuition

- Observation (e.g. in Twitter)
  - A single message contains little context
  - Noisy context (informal language)
- Intuition 1: **Spatiotemporal signals**
  - Entity prior changes over time or space
  - Location: different countries or regions
  - Time: certain duration, ppl talk certain event
- Intuition 2: **Easier surface forms**
  - Inter-tweet interactions

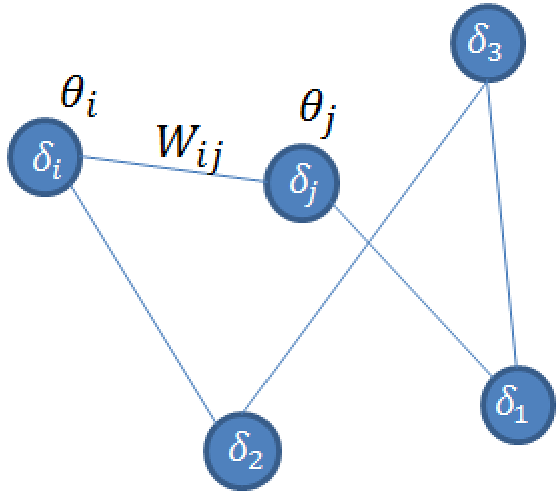
"spurs" → SA Spurs
91% in US vs. 8% in UK

"Hillary Clinton" vs. "Clinton"

Graph-based smoothing

Graph-based smoothing

- Synthetic graph (We build a graph)
  - Each node is a location or time bin
  - W_{ij} indicates the similarity between bin i and bin j (e.g. based on distance)
  - Each node has entity distribution (Clinton: Hillary 80%, bill 20% for now; in 1990 times, maybe bill 90%, Hillary 10%; if dates are very near, then their entity distributions should be same; Should time difference, e.g. 6 months, the entity should be quite similar; in the current news today and yesterday are similar)
  - Entity distribution can propagate on such our graph



Y. Fang and M.-W. Chang. Entity Linking on Microblogs with Spatial and Temporal Signals. In *TACL* 2(Oct), 2014, pp. 259−272.

Application 2:

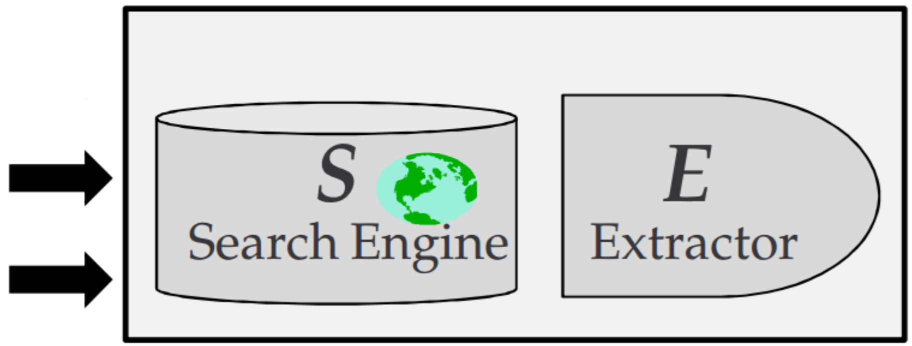
Pattern-based relation extraction

What is it?

Input: ***Seed Tuples***

$T_0 =$
{(Ottawa, Canada),
(Beijing, China)}

Corpus *D*



Output: ***Tuples***

$R =$
{(Paris, France),
(Berlin, Germany),
(Tokyo, Japan),
(Canberra, Australia)}
⋮

Insight: Pattern-tuple duality

- Tuples co-occur with text patterns

… **Beijing**, the capital of **China**, is …
… **Japan**'s capital is **Tokyo** …
… the largest city of **USA** is **NYC** …

- Duality of pattern and tuple

Beijing, China ⟺ 

⟺ ⟨city⟩, the capital of ⟨country⟩

NYC, USA ⟺ Search Engine ⟺ the largest city of ⟨country⟩ is ⟨city

Good tuples ⟺ Good patterns

Challenges

#1. What **qualities** are considered "good"?

#2. How does "goodness" **mutually reinforce**?

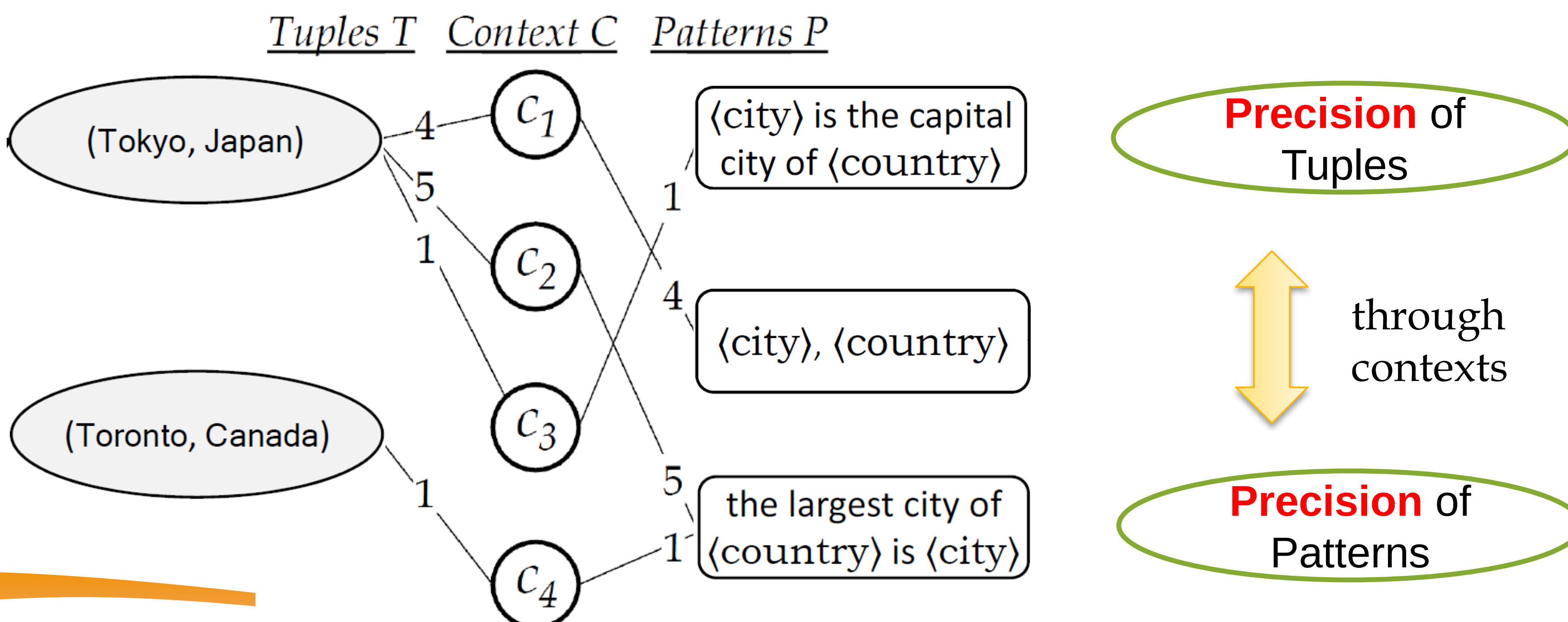
Solution to Challenge #1:
Precision and Recall

- Extraction fundamentally seeks to optimize
  - **Precision** & **Recall**
- Assessing patterns with precision & recall
  - More interpretable & better extractions

⟨city⟩ is the capital city of ⟨country⟩	⟨city⟩, ⟨country⟩
high precision low recall	low precision high recall

Solution to Challenge #2:
Syntactic Co-occurrence Graph

- Tuples and patterns co-occur
- Each co-occurrence form a *context*



References

- G. Agarwal, G. Kabra, and K. C.-C. Chang. Towards rich query interpretation: Walking back and forth for mining query templates. In *WWW* 2010, pp. 1–10.
- **X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML* 2003, pp. 912—919.**
- S. Guo, M.-W. Chang, and E. Kıcıman. 2013. To link or not to link? A study on end-to-end tweet entity linking. In *NAACL*, pp. 1020–1030.
- **Y. Fang and M.-W. Chang. Entity Linking on Microblogs with Spatial and Temporal Signals. In *TACL* 2(Oct), 2014, pp. 259–272.**
- Y. Fang, Kevin C.-C. Chang and Hady W. Lauw. Graph-based Semi-supervised Learning: Realizing Pointwise Smoothness Probabilistically. In *ICML* 2014 (2), pp. 406--414.
- Y. Fang, K. C.-C. Chang and H. W. Lauw. RoundTripRank: Graph-based Proximity with Importance and Specificity. In *ICDE* 2013, pp. 613--624.
- Y. Fang, P. Hsu and K. C.-C. Chang. Confidence-Aware Graph Regularization with Heterogeneous Pairwise Features. In *SIGIR* 2012, pp. 951--960.
- **Y. Fang and K. C.-C. Chang. Searching Patterns for Relation Extraction over the Web: Rediscovering the Pattern-Relation Duality. In *WSDM* 2011, pp. 825--834.**

Thank You

Contact: xlli@i2r.a-star.edu.sg if you have questions