Maximum Entropy Approach to Time Aware Link Prediction

Tomasz Tylenda

October 31, 2008

The problem

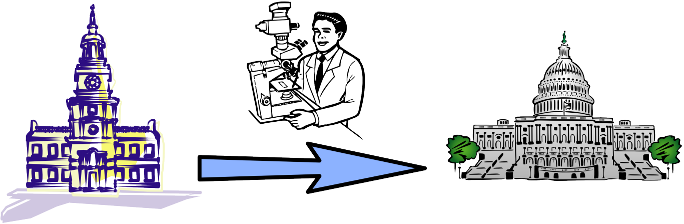
Link prediction problem occurs in many domains:

- collaboration among scientists — predict which pairs of authors are likely to collaborate in future,
- friendship — suggest new friendships on a social networking website.

A close problem is inference of missing links:

- terrorist networks — detection of unobserved connections between terrorists,
- protein interactions — suggest unobserved reactions of proteins.

Motivation for Time Awareness

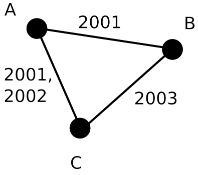


- Suppose a scientist moves to a new university.
- He will more likely work with his new colleagues than with the colleagues from the old place.
- How can we exploit this observation in link prediction?

Example

- Three authors: A, B, C.
- A — works at University 1,
- B — works at University 2,
- C in 2003 moves from University 1 to University 2.

paper id	year	A	B	C
1	2001	x	x	
2	2001	x		x
3	2002	x		x
4	2003		x	x



- How likely is that C will collaborate with B?
- How likely is that C will collaborate with A?
- C has more papers with A, but he works with B now.

Probabilistic Model

- A paper is denoted by a bit string $b_A b_B b_C$, e.g.
  110 — a paper written by A and B
  000 — other authors
- We want to construct a probability distribution $P(b_A b_B b_C)$.
- The data gives us marginals of $P(b_A b_B b_C)$ — we should respect them.
- $P(b_A b_B b_C)$ should be as close to a prior $Q(b_A b_B b_C)$ as possible (smooth and unique).

Probabilistic Model

Suppose that there are more authors. They wrote 36 additional papers. The data gives us marginal distributions:

$$P(AB) = P(11\_) = P(110) + P(111) = 1/40$$
$$P(AC) = P(1\_1) = P(101) + P(111) = 2/40$$
$$P(BC) = P(\_11) = P(011) + P(111) = 1/40$$

A	B	C
x	x	
x		x
x		x
	x	x

In general j-th constraint can be written as

$$\sum_x P(x)k(x|j) = d_j$$

where $k(event|constr.)$ is an indicator function.

Probabilistic Model

We measure distance between distributions with Kullback–Leibler divergence:

$$D(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

Optimization Problems

Plain model (Wang, et. al., ICDM 2007):

$$\min_{P} D(P||Q)$$

$$\text{s.t} \sum_{x} P(x)k(x|j) = d_j$$

$$\sum_{x} P(x) = 1$$

Time aware (our approach):

$$\min_{P,\beta} D(P||Q) + \sum_{j} w_j \beta_j$$

$$\text{s.t} \left| \sum_{x} P(x)k(x|j) - d_j \right| \le \beta_j$$

$$\beta_j \ge 0$$

$$\sum_{x} P(x) = 1$$

Choice of weights:

- Relax constraints that pertain to old events.
- Keep constraints that pertain to frequent events.

Optimization Problems

Lagrangian multipliers reveal connection between the two problems — they are almost identical.

$$P(x) = Q(x) \exp(\lambda_0) \exp\left(-1 - \sum_j \lambda_j k(x|j)\right)$$

$$g(\lambda) = \sum_x Q(x) \exp(\lambda_0) \exp\left(-1 - \sum_j \lambda_j k(x|j)\right) + \sum_j \lambda_j d_j$$

Plain (time unaware):

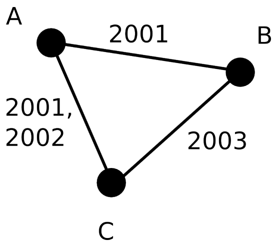
$$\lambda_0 \in \mathbb{R}$$
$$\lambda_j \in \mathbb{R}$$

Time aware:

$$\lambda_0 \in \mathbb{R}$$
$$\lambda_j \in [-w_j; w_j]$$

- λ_j are shrunk, $P(x)$ cannot attain high values,
- one implementation can solve both problems.

Numerical Experiment



Plain maxent:

$$P(AC) = 0.050$$
$$P(BC) = 0.024$$
$$P(BC)/P(AC) = 0.473$$

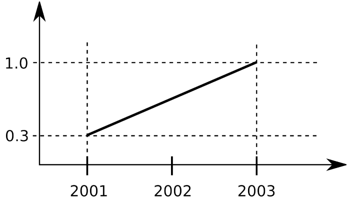
Time aware:

$$P(AC) = 0.023$$
$$P(BC) = 0.013$$
$$P(BC)/P(AC) = 0.579$$

Choice of weights

This is an open problem. Our current scheme is as follows:

- Papers are weighted in linear fashion:



- Initial weight of a constraint is the sum of weights of papers which give this constraint.

paper id	year	A	B	C	weight
1	2001	x	x		w_{2001}
2	2001	x		x	w_{2001}
3	2002	x		x	w_{2002}
4	2003		x	x	w_{2003}

$w_{AC} = w_{2001} + w_{2002}$ $\qquad$ $w_{BC} = w_{2003}$

Choice of weights

- They are subsequently divided by the maximal weight.

$$w_j \leftarrow \frac{w_j}{\max_j w_j}$$

- In order to be effective they must be multiplied by λ_j from plain maxent

$$w_j \leftarrow w_j \lambda_j$$

Choice of weights

In presence of prolific authors the weighting scheme demonstrates unwanted behaviour:

- Suppose that we have a constraint

$$P(A) = d_A$$

and A is a prolific authors with 100 papers.
- Other constraints come from average authors.
- The constraint for $P(A) = d_A$ will be assigned weight 1.0, and the rest will get very small weights.

Classifier — Design

- pairs of authors $\longrightarrow$ two classes,
  - positive pairs: pairs of authors who collaborated in the last year, but they had never collaborated before,
  - negative pairs: did not collaborate at all, **chosen at random**
- logistic regression,
- $P(BC)$ — feature for the classifier,
- other features
  - measures of the graph structure,
  - attributes of nodes.

Classifier — Testing

- Testing method for the time aware classifier is still an open problem.
- How likely is that a pair of authors will write a paper together?

type of a pair	plain	time-aware
1. no previous collab.	−	−
2. collab. long time ago	+	+
3. recent collab.	+	++

- Current testing procedure treats cases 2 and 3 as one class (positive).
- How to verify that a classifier distinguishes cases 2 and 3?
- In some scenarios case 1 is not interesting.

Bibliography

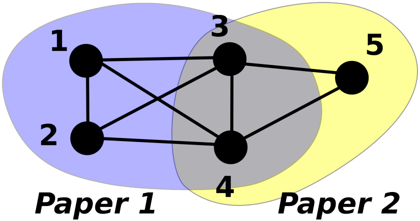
- C. Wang, V. Satuluri, S. Parthasarathy. *Local Probabilistic Models for Link Prediction.* ICDM 2007.
- D. Liben-Nowell, J. Kleinberg. *The link-prediction problem for social networks.* CIKM 2003.
- C. Faloutsos, K.S. McCurley, A. Tomkins. *Fast discovery of connection subgraphs.* KDD 2004.
- H. Tong, C. Faloutsos. *Center-piece subgraphs: problem definition and fast solutions.* KDD 2006.
- J. O'Madadhain, J. Hutchins, P. Smyth. *Prediction and ranking algorithms for event-based network data.* SIGKDD Explor. Newsl. 2005.
- F. Jelinek. *Statistical Methods for Speech Recognition.* MIT Press 1998.
- S. Boyd, L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

Sources of Information

There are three sources of information:

- event log,
- graph,
- nodes' attributes.

Paper	Author 1	2	3	4	5
1	1	1	1	1	0
2	0	0	1	1	1



Nodes' attributes — words in the titles of each author, stop words removed.

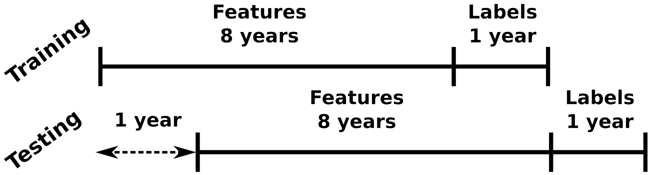
Dataset

Dataset: DBLP collection of computer science articles

- 10 years,
- 23000 authors, 18600 papers,
- 57000 collaborations.

Training and Testing

- Standard logistic regression is used for classification.
- How to obtain features and labels?



- Positive pairs: pairs of authors who collaborated in the last year, but they had never collaborated before.
- Negative pairs: did not collaborate at all, chosen at random.

Features

Three kinds of features:

- semantic,
- topological,
- co-occurrence probability.

How to use semantic features?

1. Represented the words from titles in TFIDF space,
2. the semantic feature is cosine of the angle between two vectors.

Topological Features

Topological features exploit only the structure of the graph.

- *Katz* measure:
  - $Katz(s,t) := \sum_{i=1}^{\infty} \beta^i p_i$
    p_i — number of paths of length i
    $\beta_i \in (0,1)$
  - very effective for link prediction,
  - approximation is used ($\infty \approx 4$).
- Adamic-Adar measure:
  - let $\Gamma(x)$ be the set of neighbors of vertex x,
  - $score(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} 1/\log \|\Gamma(z)\|$.

Determing nodes' neighborhood

- collect nodes on path of length 2, 3, and so on,
- paths of the same length are ordered by *frequency score* (more information on the path),
- initially the neighborhood of nodes s and t are the nodes themselves,
- paths are added until the neighborhood grows to required size.

Prior for MaxEnt

From our example we can estimate

$$P(b_A = 1),\ P(b_B = 1),\ P(b_C = 1)$$

The prior is based on independence assumption

$$Q(A, B, C) = P(A) \cdot P(B) \cdot P(C)$$

Connection Subgraphs

Faloutsos, et. al. studied the following problem:

Problem

Given a large graph G, vertices s and t, find a small (at most b nodes) subgraph that best captures the relationship between s and t.

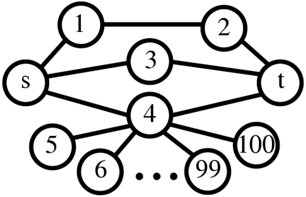
Earlier Work (1)

Surprisingly, no work discussing "good" subgraphs was found.

The standard graph-theoretic measures of goodness are

- shortest distance
- maximum flow

Earlier Work (2)



Let us consider equal weights for all edges, and vertices s and t.

Problems:

- shortest paths — “famous” nodes
- maximum flow — does not capture path’s length ($s \to 1 \to 2 \to t$ and $s \to 3 \to t$ have the same goodness)

New Approach

A new measure of goodness

- inspired by electrical currents in a network of resistors,
- edge weight corresponds to resistor conductance,
- the best subgraph can deliver the most current.

Electricity and Random Walks

Voltage $+1$ is applied to the source s, and voltage 0 is applied to the sink t. Remaining voltages and currents in the network are uniquely determined by

- Ohm's law: $I(u,v) = C(u,v)(V(u) - V(v))$
- Kirchhoff's law: $\sum_u I(u,v) = 0 \quad (u \neq s, t)$

There is a connection between currents in the electrical network and random walks on the graph.

Let us consider random walks starting from t, ending on s and following edges with probability proportional to conductance. The current $I(u,v)$ is proportional to number of times the edge (u,v) is traversed.

Universal Sink

- This method does not solve the problem with "famous nodes" and long paths.
- In order to solve it a universal sink z is introduced. It is grounded $V(z) := 0$ and connected to every node in the network.
- The connection to random walks carries through with small modifications.

Algorithms

1. Find candidate graph
   - initialize graph to be empty,
   - add paths which maximize flow divided by number of new vertices that must be added,
   - it can be computed with dynamic programming.
2. Find optimal subgraph
   - takes s and t and grow neighborhood around them
   - start with $\{s, t\}$ and add nodes
     - close to the source and the sink
     - with strong connections
     - low degree to avoid "famous node effect"