A Theoretical Justification of Link Prediction Heuristics

Deepayan Chakrabarti (deepay@cs.cmu.edu)

Purnamrita Sarkar

Andrew Moore
Link Prediction

- Which pair of nodes \( \{i,j\} \) should be connected?

Goal: Recommend a movie
Link Prediction

Which pair of nodes \{i,j\} should be connected?

Goal: Suggest friends
Link Prediction Heuristics

- Predict link between nodes
  - Connected by the shortest path
  - With the most common neighbors (length 2 paths)
  - More weight to low-degree common neighbors (Adamic/Adar)

Alice

Bob

Charlie

1000 followers

Prolific common friends

3 followers

Less evidence

Less prolific

Much more evidence
Link Prediction Heuristics

- Predict link between nodes
  - Connected by the shortest path
  - With the most common neighbors (length 2 paths)
  - More weight to low-degree common nbrs (Adamic/Adar)
  - With more short paths (e.g. length 3 paths)
    - exponentially decaying weights to longer paths (Katz measure)
  - ...

Previous Empirical Studies*

How do we justify these observations? Especially if the graph is sparse

Link prediction accuracy*

Random  Shortest Path  Common Neighbors  Adamic/Adar  Ensemble of short paths

*Liben-Nowell & Kleinberg, 2003; Brand, 2005; Sarkar & Moore, 2007
Link Prediction – Generative Model

Model:
1. Nodes are uniformly distributed points in a latent space
2. This space has a distance metric
3. Points close to each other are likely to be connected in the graph
   - Logistic distance function (Raftery+ / 2002)
Link Prediction – Generative Model

Model:
1. Nodes are uniformly distributed points in a latent space
2. This space has a distance metric
3. Points close to each other are likely to be connected in the graph

α determines the steepness
Link Prediction – Generative Model

Higher probability of linking

$\alpha$ determines the steepness

radius $r$

Model:
1. Nodes are uniformly distributed points in a latent space
2. This space has a distance metric
3. Points close to each other are likely to be connected in the graph
Link Prediction – Generative Model

Model:
1. Nodes are uniformly distributed points in a latent space
2. This space has a distance metric
3. Points close to each other are likely to be connected in the graph

Link prediction ≈ find nearest neighbor who is not currently linked to the node.

- Equivalent to inferring distances in the latent space
Previous Empirical Studies*

*Liben-Nowell & Kleinberg, 2003; Brand, 2005; Sarkar & Moore, 2007

Especially if the graph is sparse

![Bar graph showing link prediction accuracy for different methods: Random, Shortest Path, Common Neighbors, Adamic/Adar, Ensemble of short paths.](chart_image)

- Random Path
- Shortest Path
- Common Neighbors
- Adamic/Adar
- Ensemble of short paths
Common Neighbors

- $Pr_2(i,j) = Pr(\text{common neighbor}|d_{ij})$

$$Pr_2(i, j) = \int Pr(i \sim k \mid d_{ik}) Pr(j \sim k \mid d_{jk}) P(d_{ik}, d_{jk} \mid d_{ij}) \partial d_{ik} \partial d_{jk}$$

Product of two logistic probabilities, integrated over a volume determined by $d_{ij}$

As $\alpha \to \infty$, Logistic $\to$ Step function

Much easier to analyze!
Common Neighbors

- Unit volume universe
- Everyone has same radius \( r \)
- \( \text{Pr}_2(i, j) = A(r, r, d_{ij}) \)
- \( \eta = \text{Number of common neighbors} \)
- \( V(r) = \text{volume of radius } r \text{ in } D \text{ dims} \)

Distance bound:

\[
2r \left[ 1 - \left( \frac{\eta/N + \varepsilon}{V(r)} \right)^{1/D} \right] \leq d_{ij} \leq 2r \sqrt{1 - \left( \frac{\eta/N - \varepsilon}{V(r)} \right)^{2/D}}
\]
Common Neighbors

- OPT = node closest to i
- MAX = node with max common neighbors with i

Theorem: \( d_{OPT} \leq d_{MAX} \leq d_{OPT} + 2[\varepsilon/V(1)]^{1/D} \)

Link prediction by common neighbors is asymptotically optimal
Common Neighbors: Distinct Radii

- Node $k$ has radius $r_k$.
  - $i \rightarrow k$ if $d_{ik} \leq r_k$ (Directed graph)
  - $r_k$ captures popularity of node $k$

**Type 1:** $i \leftarrow k \rightarrow j$

**Type 2:** $i \rightarrow k \leftarrow j$

$A(r_i, r_j, d_{ij})$

$A(r_k, r_k, d_{ij})$
Type 2 common neighbors

Example graph:

- $N_1$ nodes of radius $r_1$ and $N_2$ nodes of radius $r_2$
- $r_1 << r_2$

\[
\eta_1 \sim \text{Bin}[N_1, A(r_1, r_1, d_{ij})] \quad \quad \eta_2 \sim \text{Bin}[N_2, A(r_2, r_2, d_{ij})]
\]

Pick $d^*$ to maximize $\Pr[\eta_1, \eta_2 | d_{ij}]$

\[
\Rightarrow w(r_1)E[\eta_1 | d^*] + w(r_2)E[\eta_2 | d^*] = w(r_1)\eta_1 + w(r_2)\eta_2
\]

Inversely related to $d^*$  
Weighted common neighbors
Common Neighbors: Distinct Radii

- Node $k$ has radius $r_k$.
- $i \rightarrow k$ if $d_{ik} \leq r_k$ (Directed graph)
  - $r_k$ captures popularity of node $k$

- “Weighted” common neighbors:
  - Predict $(i,j)$ pairs with highest $\sum w(r)\eta(r)$

$\sum w(r)\eta(r)$

Weight for nodes of radius $r$

# common neighbors of radius $r$
Type 2 common neighbors

Presence of common neighbor is very informative.

Absence is very informative.

Real world graphs generally fall in this range.

$w(r) \approx \frac{\text{const}}{r} \approx \frac{\text{const}}{\deg^{1/D}}$

$r$ is close to max radius.
Type 2 common neighbors

Real world graphs generally fall in this range

\[ w(r) \approx \frac{\text{const}}{r} \approx \frac{\text{const}}{\deg^{1/D}} \]

\( r \) is close to max radius
Previous Empirical Studies*

Especially if the graph is sparse

*Liben-Nowell & Kleinberg, 2003; Brand, 2005; Sarkar & Moore, 2007
**$\ell$ hop Paths**

- Common neighbors = 2 hop paths

**Analysis of longer paths: two components**

1. Bounding $E(\eta_\ell \mid d_{ij})$. [$\eta_\ell = \# \ell$ hop paths]
   - Bounds $\Pr_\ell(i,j)$ by using triangle inequality on a series of common neighbor probabilities.

2. $\eta_\ell \approx E(\eta_\ell \mid d_{ij})$

---

**Triangulation**
**$\ell$ hop Paths**

- Common neighbors = 2 hop paths

- Analysis of longer paths: two components
  1. Bounding $E(\eta_\ell \mid d_{ij})$. [$\eta_\ell = \# \ell$ hop paths]
    - Bounds $Pr_\ell(i,j)$ by using triangle inequality on a series of common neighbor probabilities.
  2. $\eta_\ell \approx E(\eta_\ell \mid d_{ij})$
    - Bounded dependence of $\eta_\ell$ on position of each node
      - Can use McDiarmid’s inequality to bound
        $$|\eta_\ell - E(\eta_\ell \mid d_{ij})|$$
ℓ-hop Paths

- Common neighbors = 2 hop paths

- For longer paths:
  \[ d_{ij} \leq r + (\ell - 1)r[1 - g(\eta_\ell, N, \delta)] \]

- Bounds are weaker

- For \( \ell' \geq \ell \) we need \( \eta_{\ell'} \gg \eta_\ell \) to obtain similar bounds
  - \( \Rightarrow \) justifies the exponentially decaying weight given to longer paths by the Katz measure
Summary

Three key ingredients

1. Closer points are likelier to be linked.

2. Triangle inequality holds
   ➔ necessary to extend to ℓ-hop paths

3. Points are spread uniformly at random
   ➔ Otherwise properties will depend on location as well as distance
The number of paths matters, not the length.

*Liben-Nowell & Kleinberg, 2003; Brand, 2005; Sarkar & Moore, 2007
Summary

For large dense graphs, common neighbors are enough.

The number of paths matters, not the length.

*Liben-Nowell & Kleinberg, 2003; Brand, 2005; Sarkar & Moore, 2007*
Differentiating between different degrees is important.

For large dense graphs, common neighbors are enough.

The number of paths matters, not the length.

*Liben-Nowell & Kleinberg, 2003; Brand, 2005; Sarkar & Moore, 2007
In sparse graphs, paths of length 3 or more help in prediction.

For large dense graphs, common neighbors are enough.

The number of paths matters, not the length.

Differentiating between different degrees is important.

*Liben-Nowell & Kleinberg, 2003; Brand, 2005; Sarkar & Moore, 2007
Sweep Estimators

$Q_r = \text{Fraction of nodes with radius } \leq r \text{ which are common neighbors}$

Large $Q_r \Rightarrow$ small $d_{ij}$

$T_R = \text{Fraction of nodes with radius } \geq R \text{ which are common neighbors}$

Small $T_R \Rightarrow$ large $d_{ij}$

Number of common neighbors of a given radius $r$