CS 768: Learning With Graphs **Autumn 2020-2021**

Lecture 8: Theoretical Justification of Link Prediction Heuristic

Instructor: Prof. Abir De *Scribe: Pintu Kumar, Prantik Chakraborty*

Let G be the orignal graph and $]\hat{G}$ be a graph generated by some link prediction algorithm. Then we want to see a measure of $\Delta(\hat{G}, G)$.

8.1 Generative Model of G

- Build a hyper sphere of volume 1 in the space $\mathbb{R}^d$.
- Distribute all the nodes within the sphere uniformly at random.
- Set $(u, v) \in E$ if $d_{uv} < r$.

8.2 Oracle Algorithm A*

Suppose a Link prediction algorithm A^* knows distance between any pair of nodes. Then sorting edge in basis of d_{uv} gives a algorithm with average precision 1.
Let A be any other LP algorithm, If the ranked list given by A is close to the ranked list generated by A^* then we can say G is close to $\hat{G}$.
This is difficult to prove so we prove $d_{min} \approx d_{uv}$, where u, v has highest LP score. This is similar to proving $||d_{min} - d_{uv}|| < \epsilon$ with high probability.

8.2.1 Key Points

1. An oracle algorithm which has access to the latent distance between the nodes u and v
2. In practice the link prediction algorithm does not have access to the latent distance between the nodes u and v. It can only detect the absence or presence of an edge
3. There is a generative process using which the graphs are created

8.3 Approach

Main idea: We assume we can approximate the latent distance between u and v from the graph and the create a ranked list

To prove: Adamic-Adar and Common-Neighbor can give a reasonable approximate for the latent distance between u and v. The ranked list given by the oracle algorithm and the Adamic-Adar/Common-Neighbor would be roughly same.

8.4 Common Neighbours

Let G=$\{V, E\}$ be a graph and $N(u)$ denote set of neigbours of $u \in V$. Then for $u, v \in V$

$$\begin{aligned} CN(u,v) &= |N(u) \cap N(v)| \\ &= \sum_{w \in V} |\mathbb{1}\{w \in N(u) \cap N(v)\}| \end{aligned} \tag{8.1}$$

In order to find notation of common neighbours in generated model, calculate $\mathbb{E}[CN(u,v)]$
$\mathbb{E}[CN(u,v)] = \sum_{w \in V} \mathbb{P}(w \in N(u) \cap N(v))$.
To calculate the probability inside summation, we first calculate $\mathbb{P}(w \in N(u) \cap N(v) | d_{uv})$.

$$\begin{aligned} \mathbb{P}(w \in N(u) \cap N(v) | d_{uv}) &= \mathbb{P}(u - v - w | d_{uv} \\ &= \int_{d_{uw}, d_{wv}} \mathbb{P}(u - w | d_{uw}).\mathbb{P}(w - v | d_{wv}).\mathbb{P}(d_{uw}, d_{wv} | d_{uv}) \\ &= A(r, d_{uv}) \end{aligned} \tag{8.2}$$

Here $u - v$ means edge exist between nodes u and v. See 8.1 for $A(r, d_{uv})$

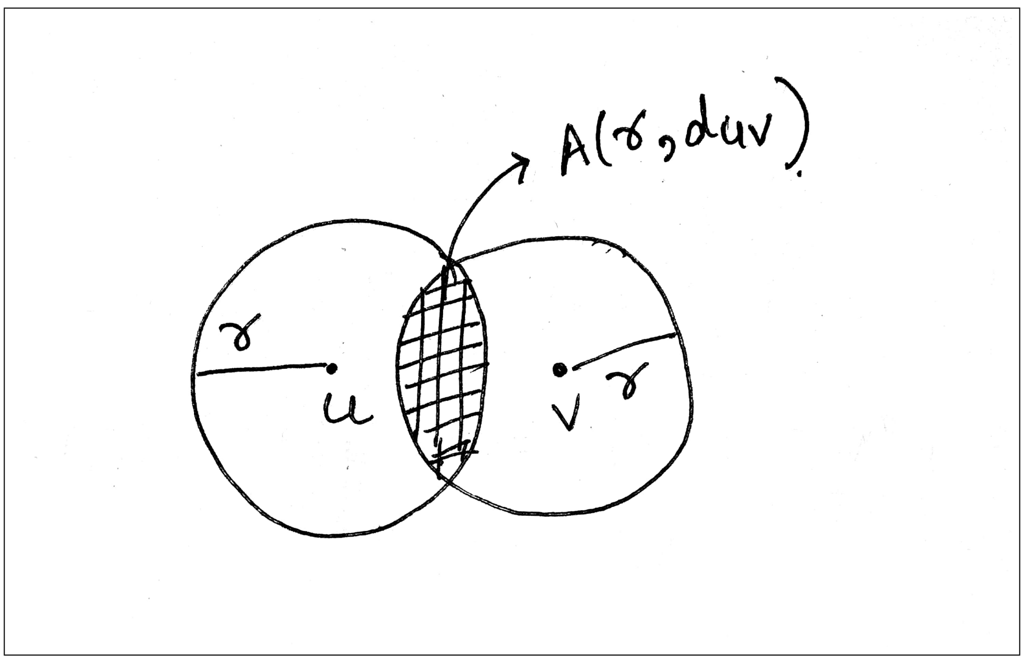

Figure 8.1: Pictorial representation of $A(r, d_{uv})$

Then

$$\begin{aligned} \mathbb{E}[CN(u,v)] &= \sum_{w \in V} \int A(r, d_{uv}).\mathbb{P}(d_{uv}) \\ &= |V| \int A(r, d_{uv}).\mathbb{P}(d_{uv}) \end{aligned} \tag{8.3}$$

This is difficult to compute so variance and expectation of $\mathbb{1}(w \in N(u) \cap N(v) | d_{uv})$ is used for the proof.

Let X_w denote $\mathbb{1}(w \in N(u) \cap N(v) | d_{uv})$. Then concentration inequality gives.

$$\mathbb{P}(|\sum_{w \in V} \frac{X_w}{|V|} - \mathbb{E}[X_w]| > \epsilon) \leq 2\delta, \tag{8.4}$$

where

$$\epsilon = \sqrt{\frac{2Var(X_w).log(2/\delta)}{|V|}} + \frac{3.5log(2/\delta)}{3(|V|-1|)} \tag{8.5}$$

Using inequality 8.4 we show that distance d_{uv} between the node u and v corresponds to the ones giving highest common neighbours score is close to the least possible distance d_{min} (corresponding to A^*)

Rather than showing $d_{uv} \approx d_{min}$, we show $A(r, d_{uv}) \approx A(r, d_{min})$. Showing this implies $d_{uv} \approx d_{min}$ asymptotically with size of graph.

$$\begin{aligned} \mathbb{E}[X_w] &= A(r, d_{uv}) \\ &= \frac{2\pi^{\frac{D-1}{2}} r^D}{\Gamma((D-1)/2)} \int_0^{cos^{-1}(\frac{d_{uv}}{2r})} sin^D(t)dt \end{aligned} \tag{8.6}$$

A bound on $A(r, d_{uv})$ is given by,

$$\left(1 - \frac{d_{uv}}{2r}\right)^D * v(r) \leq A(r, d_{uv}) \leq \left(1 - \left(\frac{d_{uv}}{2r}\right)^2\right)^{\frac{D}{2}} * v(r) \tag{8.7}$$

Let us define few notation

$$\epsilon_0 = \sqrt{\frac{2Var(X_w^*).log(2/\delta)}{|V|}} + \frac{3.5log(2/\delta)}{3(|V|-1|)} \tag{8.8}$$

X_w^* : Optimum given by A^*

$$\epsilon_m = \sqrt{\frac{2Var(X_w^{CN}).log(2/\delta)}{|V|}} + \frac{3.5log(2/\delta)}{3(|V|-1|)} \tag{8.9}$$

X_w^{CN} : Optimum given by common neighbours. Then following inequality can be obtained,

$$\mathbb{P}(A(r, d_{uv}) > A(r, d_{min}) - \epsilon_0 - \epsilon_m) \geq 1 - 2\delta \tag{8.10}$$

Here d_{uv} is the distance between nodes u and v which gives best score for common neighbours, $d_{,min}$ is the distance between nodes u^* and v^* which gives best score for A^*,