Theoretical Justification of Popular Link Prediction Heuristics

Purnamrita Sarkar
Carnegie Mellon University
psarkar@cs.cmu.edu

Deepayan Chakrabarti
Yahoo! Research
deepay@yahoo-inc.com

Andrew W. Moore
Google, Pittsburgh
awm@google.com

Abstract

There are common intuitions about how social graphs are generated (for example, it is common to talk informally about *nearby* nodes sharing a link). There are also common heuristics for predicting whether two currently unlinked nodes in a graph should be linked (e.g. for suggesting friends in an online social network or movies to customers in a recommendation network). This paper provides what we believe to be the first formal connection between these intuitions and these heuristics. We look at a familiar class of graph generation models in which nodes are associated with locations in a latent metric space and connections are more likely between closer nodes. We also look at popular link-prediction heuristics such as number-of-common-neighbors and its weighted variants (Adamic & Adar, 2003) which have proved successful in predicting missing links, but are not direct derivatives of latent space graph models. We provide theoretical justifications for the success of some measures as compared to others, as reported in previous empirical studies. In particular we present a sequence of formal results that show bounds related to the role that a node's degree plays in its usefulness for link prediction, the relative importance of short paths versus long paths, and the effects of increasing non-determinism in the link generation process on link prediction quality. Our results can be generalized to any model as long as the latent space assumption holds.

1 Introduction

Link prediction is a key problem in graph mining. It underlies recommendation systems (e.g., movie recommendations in Netflix, music recommendation engines like `last.fm`), friend-suggestions in social networks, market analysis, and so on. As such, it has attracted a lot of attention in recent years, and several heuristics for link prediction have been proposed (Adamic & Adar, 2003). In-depth empirical studies comparing these heuristics have also been conducted (Liben-Nowell & Kleinberg, 2003) and (Brand, 2005), and two observations are made consistently: (1) a simple heuristic, viz., predicting links between pairs of nodes with the most common neighbors, often outperforms more complicated heuristics, (2) a variant of this heuristic that weights common neighbors using a carefully chosen function of their degrees (Adamic & Adar, 2003) performs even better on many graphs and (3) heuristics which use an ensemble of short paths between two nodes (Katz, 1953) often perform better than those which use longer paths. However, there has been little theoretical work on why this should be so. We present, to our knowledge, the first theoretical analysis of link prediction on graphs. We show how various heuristics compare against each other, and under what conditions would one heuristic be expected to outperform another. We are able to provide theoretical justifications for all of the empirical observations mentioned above.

We define the link prediction problem as follows. There is a latent space in which the nodes reside, and links are formed based on the (unknown) distances between nodes in this latent space. Individual differences between nodes can also be modeled with extra parameters. The quality of link prediction now depends on the quality of estimation of distance between points. We show how different estimators provide *bounds* on distance. Clearly, the tighter the bounds, the better we can distinguish between pairs of nodes, and thus the better the quality of link prediction.

While any latent space model can be used, we extend a model by (Raftery et al., 2002) due to two characteristics: (1) it is simple to state and analyze, (2) yet, it is powerful enough to show all of the effects that affect estimation, such as node degree, lengths of paths, etc. Our results do not assume any degree distribution on the graph; in fact, they depend on very simple properties that should be generalizable to other models as well.

Our primary contributions are as follows:

- We formulate the link prediction problem as a problem of estimating distances between pairs of nodes, where the nodes lie at unknown positions in some latent space and the observed presence or absence of links between nodes provides clues about their distances.

- We show that the number of common neighbors between a pair of nodes gives bounds on the distance between them, with the upper bound on distance decreasing quickly as the count of common neighbors increases. This justifies the popular heuristic of predicting links simply by picking node pairs with the maximum number of common neighbors.
- Empirical studies (Liben-Nowell & Kleinberg, 2003) have shown that another popular heuristic (Adamic & Adar, 2003) that uses a carefully *weighted* count of common neighbors often outperforms the unweighted count. We present theoretical justification for this, and generalize it to other possible weighting schemes that should be similarly useful.
- Finally, another set of heuristics consider longer paths between pairs of nodes, e.g., hitting-time and other measures based on random walks. Our results here are twofold. (1) We show that while the number of long paths can, indeed, provide bounds on distance, these are looser than the bounds obtained if enough short paths (or ideally, common neighbors) exist. Thus, longer paths are more useful if shorter paths are rare or non-existent. (2) We also show that the bounds obtained from long paths can get much tighter given just the knowledge of *existence* of a short path. Thus, even the existence of a single short path can improve bounds obtained from long paths.
- Our results can be applied to any social network model where: nodes are distributed independently in some latent metric space; probability of a link satisfies *homophily*; given the positions links are independent of each other.

This paper is organized as follows. In section 2 we introduce related work and background on link prediction and latent space models. Sections 3 and 4 consist of a formal relationship between popular heuristics like common neighbors and our graph model with same and distinct radii. In section 5, we analyze the implication of paths of length $\ell > 2$. Section 6 shows how to extend the analysis to handle non-determinism in the link generation process. In section 7, we summarize this paper and discuss several implications of our work.

2 Review of Previous Empirical Studies

We will briefly describe the link prediction problem, and the observations from previous empirical studies. Then we will describe the latent space model we use, and conclude with the relation of this model to link prediction.

2.1 Link Prediction

Many real world graph-mining problems can be framed as link prediction. Popular applications include suggesting friends on Facebook, recommending movies (Netflix, MovieLens) or music (last.fm, Pandora) to users. Link prediction on informal office-network has been shown to be useful for suggesting potential collaborators (Raghavan, 2002). Also in intelligence analysis (Schroeder et al., 2003), link-prediction can suggest potential involvement between a group of individuals, who do not have prior records of interaction.

In general popular graph-based proximity measures like personalized pagerank (Jeh & Widom, 2002), hitting and commute times (Aldous & Fill, 2001), Adamic/Adar (Adamic & Adar, 2003) are used for link prediction on graphs. The experimental setup varies from predicting all absent links at once (Liben-Nowell & Kleinberg, 2003) to predicting the best link for a given node (Brand, 2005),(Sarkar & Moore, 2007).

These papers mostly use co-authorship graphs and movie-recommendation networks. We will sketch a few of the observations from earlier empirical evaluation. We would divide heuristics into roughly two parts, simple variants of the number of common neighbors (Adamic/Adar, Jaccard etc.), and measures based on ensemble of paths. Here is a summary of the most interesting observations from (Liben-Nowell & Kleinberg, 2003; Brand, 2005), and (Sarkar & Moore, 2007).

1. The number of common neighbors performs surprisingly well on most data-sets, and in many cases beats more complex measures (Liben-Nowell & Kleinberg, 2003).

2. Adamic/Adar, which is analogous to common neighbors with a skewed weighting scheme mostly outperforms number of common neighbors (Liben-Nowell & Kleinberg, 2003).

3. Shortest path performs consistently poorly (Liben-Nowell & Kleinberg, 2003) and (Brand, 2005). We have also noted this behavior.

4. Ensemble of paths which looks at long paths (hitting and commute times) does not perform very well (Liben-Nowell & Kleinberg, 2003) and (Brand, 2005).

5. Ensemble of paths which down-weights long paths exponentially (e.g. Katz, personalized pagerank) perform better than those which are sensitive to long paths (Liben-Nowell & Kleinberg, 2003) and (Brand, 2005).

While these are interesting results, there has not been any work which theoretically justifies this behavior of different heuristics for link prediction on social networks. In our work, we analyze a simple social network model to justify these empirical results.

2.2 Latent Space Models for Social Network Analysis

Social network analysis has been an active area of research in sociology and statistics for a long time. One important assumption in social networks is the notion of homophily. (McPherson et al., 2001) write in their well-known paper, *"Similarity breeds connection. This principle-the homophily principle-structures network ties of every type, including marriage, friendship, work, advice, support, information transfer, exchange, comembership, and other types of relationship."* More formally, there is a higher probability of forming a link, if two nodes have similar characteristics. These characteristics can be thought of as different features of a node, i.e. geographic location, college/university, work place, hobbies/interests etc. This notion of *social space* has been examined by (McFarland & Brown, 1973) and (Faust, 1988).

In 2002, (Raftery et al., 2002) introduced a statistical model which explicitly associates every node with locations in a D-dimensional space; links are more likely if the entities are close in latent space. In the original model, the probability of a link between two nodes is defined as a logistic function of their distance. All the pairwise events are independent, conditioned on their latent positions, i.e. distances in the latent space. We alter this model to incorporate radius r in the exponent (for the RHH model $r = 1$). r can be interpreted as the sociability of a node. We name this model- the non-deterministic model (section 6).

$$\texttt{RHH}\text{ model:}\quad P(i \sim j|d_{ij}) = \frac{1}{1+e^{\alpha(d_{ij}-1)}} \qquad \text{Our model:}\quad P(i \sim j|d_{ij}) = \frac{1}{1+e^{\alpha(d_{ij}-r)}}$$

The model has two parameters α and r. Parameter $\alpha \geq 0$ controls the sharpness of the function whereas r determines the threshold. Setting $\alpha = \infty$ in our model, yields a simple deterministic model, on which we build our analysis. It may be noted that given the distances the links are deterministic; but given the links, inferring the distances is an interesting problem. In section 6, we show how this analysis can be carried over to the non-deterministic case with large but finite α. This assumption is reasonable, because low values of α leads to a random graph; $\alpha = 0$ is exactly the $\mathcal{G}_{N,1/2}$ random graph model. Clearly, it is impossible to perform better than a random predictor in a random graph. With distinct radii for each node, the deterministic model can be used for generating both undirected and directed graphs; however for simplicity we use a realistic directed graph model (section 4).

We assume that the nodes are uniformly distributed in a D dimensional Euclidian space. Hence $P(d_{ij} \leq x) = V(1)x^D$, where $V(1)$ is the volume of a *unit radius hypersphere*. This uniformity assumption has been made in earlier social network models, e.g. by (Kleinberg, 2000), where the points are assumed to lie on a two dimensional grid. In order to normalize the probabilities, we assume that all points lie inside a *unit volume hypersphere* in D dimensions. The maximum r satisfies $V(r) = V(1)r^D = 1$.

Connection to the Link Prediction Problem. A latent space model is well-fitted for link prediction because, for a given node i, the most likely node it would connect to is the non-neighbor at the smallest distance. The measure of distance comes from the definition of the model, which could be:

1. Undirected graph with identical r. Here distance from node i to j is simply d_{ij}.

2. Directed graph where i connects with j if d_{ij} is smaller than radius of j (or smaller that radius of i) leading to a distance of $d_{ij} - r_j$ (or $d_{ij} - r_i$).

In all these cases, the predicting distances between a pair of nodes is the key. While this can be obtained by maximizing the likelihood of the underlying statistical model, we show that one can obtain high probability bounds on distances from graph based heuristics. In fact we show that the distance to the node picked using a popular heuristic is within a small factor of the *true distance*. This factor quickly goes to zero as N becomes large. Although our analysis uses an extension of the RHH model to actually obtain the bounds on distances, the only property we use is of homophily in a latent metric space, i.e. if two nodes are *close* in some social space, then they are likely to form a link. Hence this idea should carry over to other social network models as well.

3 Deterministic Model with Identical Radii

Consider a simple version of the RHH model where all radii are equal to r, and $\alpha \to \infty$. This implies that two nodes i and j share a link (henceforth, $i \sim j$) iff the distance d_{ij} between them is constrained by $d_{ij} < r$. Thus, given node positions, links are deterministic; however the node positions are still non-deterministic. While this might appear to be a strong constraint, we will show later in Section 6 that similar results are applicable even for finite but large α. We now analyze the simplest of heuristics: counting the common neighbors of i and j. Let there be N nodes in total.

Let $\mathcal{N}(i)$ be the set of neighbors of node i. Let Y_k be a random variable which is 1 if $k \in \mathcal{N}(i) \cap \mathcal{N}(j)$, and 0 otherwise. Given d_{ij}, for all $k \notin \{i, j\}$, the Y_k's are independent since they only depend on the position

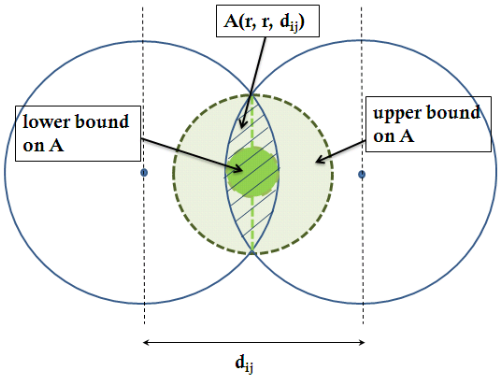

Figure 1: Common neighbors of two nodes must lie in the intersection $A(r, r, d_{ij})$.

of point k. Hence, we have

$$E[Y_k|d_{ij}] = P(i \sim k \sim j|d_{ij}) = \int\limits_{d_{ik},d_{jk}} P(i \sim k|d_{ik})P(j \sim k|d_{jk})P(d_{ik}, d_{jk}|d_{ij})d(d_{ij}) \qquad (1)$$

In the deterministic model, this quantity is exactly equal to the volume of intersection of two balls of radius r centered at i and j (see Figure 1). Denote this volume by $A(r, r, d_{ij})$. Also, the observed value of $\sum_k Y_k$ is simply the number of common neighbors η. From now on we will drop the d_{ij} part when we write expectation for notational convenience. However any expectation in terms of area of intersection is obviously computed *given the pairwise distance* d_{ij}. Thus by using empirical Bernstein bounds (Maurer & Pontil, 2009), we have:

$$P\left[\left|\sum_k Y_k/N - E[Y_k]\right| \geq \sqrt{\frac{2\mathtt{var}_N(Y)\log 2/\delta}{N}} + \frac{7\log 2/\delta}{3(N-1)}\right] \leq 2\delta \qquad (2)$$

$\mathtt{var}_N(Y)$ is the sample variance of Y, i.e. $\dfrac{\eta(1-\eta/N)}{N-1}$. Setting $\epsilon = \sqrt{\dfrac{2\mathtt{var}_N(Y)\log 2/\delta}{N}} + \dfrac{7\log 2/\delta}{3(N-1)}$

$$\boxed{P\left[\frac{\eta}{N} - \epsilon \leq A(r, r, d_{ij}) \leq \frac{\eta}{N} + \epsilon\right] \geq 1 - 2\delta} \qquad (3)$$

$A(r, r, d_{ij})$ is twice the volume of a spherical cap whose base is $d_{ij}/2$ distance away from the center:

$$A(r, r, d_{ij}) = 2\frac{\pi^{\frac{D-1}{2}} r^D}{\Gamma\left(\frac{D+1}{2}\right)} \int\limits_0^{\cos^{-1}\left(\frac{d_{ij}}{2r}\right)} sin^D(t)dt \qquad (4)$$

Given the above bounds on $A(r, r, d_{ij})$, we can obtain bounds on d_{ij} by solving eq. (4) numerically. However, weaker analytic formulas can be obtained by using hyperspheres bounding this intersection, as shown in Figure 1. $V(r)$ is the volume of a D dimensional hypersphere of radius r.

$$\left(1 - \frac{d_{ij}}{2r}\right)^D \leq \frac{A(r, r, d_{ij})}{V(r)} \leq \left(1 - \left(\frac{d_{ij}}{2r}\right)^2\right)^{D/2} \qquad (5)$$

Using this in eq. (3) gives us bounds on d_{ij}:

$$2r\left(1 - \left(\frac{\eta/N + \epsilon}{V(r)}\right)^{1/D}\right) \leq d_{ij} \leq 2r\sqrt{1 - \left(\frac{\eta/N - \epsilon}{V(r)}\right)^{2/D}}$$

Using common neighbors in link prediction. Let us recall that in link prediction, we want to pick the node which is most likely to be a neighbor of i, and is not currently a neighbor (call this OPT). If we knew the positions, we would pick the non-neighbor with the minimum distance (d_{OPT}). However, since positions in latent space are unknown, we instead predict a link to the node that shares the most common neighbors with i (call this MAX). Here we will show that (Lemma 3.2) the distance to the node with largest common neighbors (d_{MAX}) is within an additive factor of d_{OPT}. This factor goes to zero as N increases. This again shows that, as N increases, link prediction using the number of common neighbors converges to the optimal prediction.

Let the number of common neighbors between i and OPT be η_{OPT}, and between i and MAX be η_{MAX}. Now we will try to relate d_{OPT} with d_{MAX}. Note that both these distances are larger than r. Denote the area of intersections of these two nodes with i as A_{OPT} and A_{MAX} respectively. Then, we have:

Lemma 3.1 *Define* $\epsilon_o = \sqrt{\frac{2\mathtt{var}_N(Y_{OPT})\log 2/\delta}{N}} + \frac{7\log 2/\delta}{3(N-1)}$ *and* $\epsilon_m = \sqrt{\frac{2\mathtt{var}_N(Y_{MAX})\log 2/\delta}{N}} + \frac{7\log 2/\delta}{3(N-1)}$, *where Y_{OPT} and Y_{MAX} denote the random variable for common neighbors between i and OPT, and i and MAX respectively.*

$$A_{OPT} \geq A_{MAX} \qquad\qquad P\left[A_{MAX} \geq A_{OPT} - \epsilon_o - \epsilon_m\right] \geq 1 - 2\delta$$

Proof: Using the high probability bounds from eq. (3) we have (w.h.p),

$$A_{MAX} - A_{OPT} \geq \frac{\eta_{MAX}}{N} - \epsilon_m - (\frac{\eta_{OPT}}{N} + \epsilon_o)$$

By definition, $\eta_{MAX} \geq \eta_{OPT}$. This and the high probability empirical Bernstein bound on A_{OPT} yield the result. ■

This means that as N becomes large, the node with the highest number of common neighbors will be the optimal node for link prediction. Now we will give a bound on how far d_{OPT} is from d_{MAX}.

Theorem 3.2 *Define* $\epsilon_o = \sqrt{\frac{2\mathtt{var}_N(Y_{OPT})\log 2/\delta}{N}} + \frac{7\log 2/\delta}{3(N-1)}$, $\epsilon_m = \sqrt{\frac{2\mathtt{var}_N(Y_{MAX})\log 2/\delta}{N}} + \frac{7\log 2/\delta}{3(N-1)}$, *and* $\epsilon_f = \epsilon_o + \epsilon_m$.

$$d_{OPT} \leq d_{MAX} \overset{w.h.p}{\leq} d_{OPT} + 2r\left(\frac{\epsilon_f}{V(r)}\right)^{1/D} \leq d_{OPT} + 2\left(\frac{\epsilon_f}{V(1)}\right)^{1/D}$$

4 Deterministic Model with Distinct Radii

Until now our model has used the same r for all nodes. The degree of a node is distributed as $Bin(N, V(r))$, where $V(r)$ is the volume of a radius r. Thus r determines the degree of a node in the graph, and identical r will lead to a roughly regular graph. In practice, social networks are far from regular. In order to accommodate complex networks we will now allow a different radius (r_i) for node i. For this section, we will assume that these radii are given to us. The new connectivity model is: $i \rightarrow j$ iff $d_{ij} \leq r_j$, where $i \rightarrow j$ now represents a *directed* edge from i to j. While variants of this are possible, this is similar in spirit to a citation network, where a paper i tends to cite a well-cited paper j (with larger number of in-neighbors) than another infrequently cited paper on the same topic; here, r_j can be thought of as the measure of popularity of node j. Under this model, we will show why some link prediction heuristics work better than others.

As in the previous section, we can use common neighbors to estimate distance between nodes. We can count common neighbors in 4 different ways as follows:

(Type-1) All k, s.t. $k \rightarrow i$ and $k \rightarrow j$: all nodes which point to both i and j. The probability of this given d_{ij} is $P(d_{ik} \leq r_i \cap d_{jk} \leq r_j | d_{ij})$, which can be easily shown to be $A(r_i, r_j, d_{ij})$.

(Type-2) All k, s.t. $i \rightarrow k$ and $j \rightarrow k$: all nodes to which both i and j point. The probability of this given d_{ij} is $A(r_k, r_k, d_{ij})$.

(Type-3) All k, s.t. $i \rightarrow k$ and $k \rightarrow j$: all directed paths of length 2 from i to j. The probability of this given d_{ij} is given by $A(r_k, r_j, d_{ij})$.

(Type-4) All k, s.t. $j \rightarrow k$ and $k \rightarrow i$: all directed paths of length 2 from j to i. The probability of this given d_{ij} is given by $A(r_i, r_k, d_{ij})$.

If we count type-1 nearest neighbors, the argument from section 3 carries over, and if there are enough common neighbors of this type, we can estimate d_{ij} by computing $A(r_i, r_j, d_{ij})$. However, if both r_i and r_j are small, there might not be many common neighbors; indeed, if $d_{ij} > r_i + r_j$, then there will be no type-1 common neighbors. In such cases, we consider type-2 neighbors, i.e. the ones which both i and j point to. The analysis for type-3 and type-4 neighbors is very similar to that for type-2, and hence we do not discuss these any further. In the type-2 case, the radii r_k of the common neighbors play an important role. Intuitively, if both i and j point to a very popular node (high radius r_k), then that should not give us a lot of information about d_{ij}, since it is not very surprising. In particular, any type-2 common neighbor k leads to the following constraint: $d_{ij} \leq d_{ik} + d_{jk} \leq 2r_k$. Obviously, the bound is stronger for small values of r_k. This argues for weighting common neighbors differently, depending on their radii. We formalize this intuition using a toy example now. The analysis in the following section can be generalized to graphs, where the radii of the nodes form a finite set.

Recall that common neighbors can be expressed as a sum of random variables Y_k, which is 1, if k is a common neighbor of i and j.

Motivating Example. Take a toy network where the nodes can have two different radii R and R', with $R < R'$. The total number of low radii nodes is N_R, whereas that of large radii nodes is $N_{R'}$.

The expectation of the number of type-2 common neighbors will now have a mixture of $A(R, R, d_{ij})$ and $A(R', R', d_{ij})$. One solution is to estimate high probability bounds on distances from the two different classes of common neighbors separately, and then examine the intersection of these bounds. We will discuss a new estimator based on this intuition at the end of this section. The other solution is to look at weighted combinations of common neighbors from different radii. The weights will reflect how important one common neighbor is relative to another. For example, consider a pair of papers which both cite a book on introduction to algorithms (cited by 5000 other papers, i.e. higher radius), and a specific article on randomized algorithms (cited by 30 other papers, i.e. lower radius). The second article gives more evidence on the "closeness" or similarity of the pair. We will consider this approach next.

Suppose we observe η_R common neighbors of N_R nodes of small radius, and $\eta_{R'}$ common neighbors of $N_{R'}$ nodes of large radius, between pair of nodes i, j. The likelihood of these observations, given the pairwise distance d_{ij} is:

$$P(\eta_R, N_R, \eta_{R'}, N_{R'}|d_{ij}) = \prod_{r \in \{R, R'\}} \binom{N_r}{\eta_r} A(r, r, d_{ij})^{\eta_r} (1 - A(r, r, d_{ij}))^{N_r - \eta_r} \tag{6}$$

We want to rank pairs of nodes using the distance estimate d^*, which maximizes the likelihood of this partial set of observations. However, if $\eta_R > 0$, the logarithm of the above is defined only when $d_{ij} \leq 2R$. To make the likelihood well-behaved, we introduce a small noise parameter β: node i connects to node j with probability $1 - \beta$ (if $d_{ij} \leq r_j$), or with probability β (otherwise). Now, the probability of having a type-2 common neighbor of radius r will be $\beta + A(r, r, d_{ij})(1 - \beta)$. For ease of exposition we will denote this by $A_\beta(r, r, d_{ij})$. The new likelihood will be exactly as in eq. (6), except we will use A_β instead of A. Setting the derivative of the logarithm yields:

$$w(R, d^*) N_R A_\beta(R, R, d^*) + w(R', d^*) N_{R'} A_\beta(R', R', d^*) = w(R, d^*)\eta_R + w(R', d^*)\eta_{R'} \tag{7}$$

where, $w(R, d^*) = \dfrac{-\left.\frac{dA_\beta(R,R,d_{ij})}{d_{ij}}\right|_{d^*}}{A_\beta(R, R, d^*)(1 - A_\beta(R, R, d^*))}$. Note that the negative sign is only to make both sides positive, since A_β decreases with distance.

Using Leibnitz's rule on eq. 4, the derivative of A w.r.t d_{ij} can be written as:

$$\begin{aligned} A'(R, R, d_{ij}) = \frac{dA(R, R, d_{ij})}{d_{ij}} &= \begin{cases} -C_D r^{D-1} \left(1 - \frac{d_{ij}^2}{4r^2}\right)^{\frac{D-1}{2}} & \text{If } d_{ij} \leq 2r \\ 0 & \text{Otherwise} \end{cases} \\ A'_\beta(R, R, d_{ij}) = \frac{dA_\beta(R, R, d_{ij})}{d_{ij}} &= (1 - \beta) A'(R, R, d_{ij}) \end{aligned} \tag{8}$$

Suppose we could approximate $w(R, d_{ij})$ with some w_R that depends only on R and not on d_{ij}. Then, the RHS of eq. (7) can be obtained from the data alone. Also, it can be shown that the L.H.S. of eq. (8) is a monotonically decreasing function of d^*. Hence, a pair of nodes with higher RHS will have lower distance estimate, implying that ranking based on the RHS will be equivalent to ranking based on distance. All that remains is finding a good approximation w_R.

We start by bounding $A'(R, R, d_{ij})$ in terms of $A(R, R, d_{ij})$. Consider $D > 1$[1]. Combining eq. (8) with eq. (5) yields a lower bound: $A'(R, R, d) \geq c'_D \frac{A}{\sqrt{r^2 - d^2/4}}$. For the upper bound, we note that the volume of the spherical cap can be lower bounded by the volume of a sphere in $D-1$ dimensions of radius $\sqrt{r^2 - d^2/4}$,

[1] When $D = 1$, A' is constant, and $A(r, r, d) = 2r - d$.

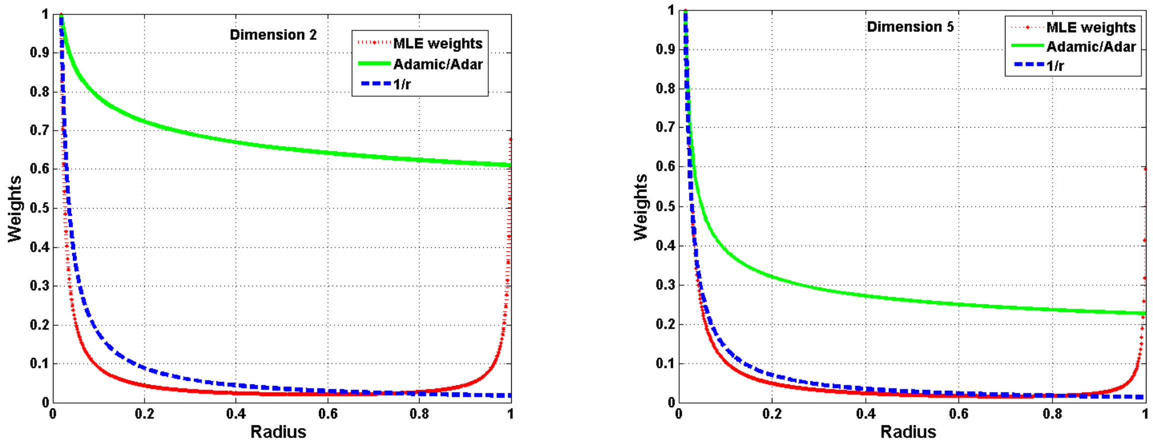

Figure 2: For a given distance d, we plot $w(r, d)$, $Adamic/Adar$ and $1/r$ with increasing r. Note that both axes are scaled to a maximum of 1. Note that $1/r$ closely matches $w(r, d)$ for a wide range of radii.

times the height $r - d/2$, times a constant depending only on D: $A(r, r, d) \geq k_D \left(r^2 - \frac{d^2}{4}\right)^{\frac{D-1}{2}} \left(r - \frac{d}{2}\right)$ Combining with eq. (8) gives: $-A'(r, r, d) \leq k'_D \frac{A}{r-d/2}$. Given that β is extremely small, we have

$$\frac{1}{r} c''_D \frac{1}{\sqrt{1 - \frac{d^2}{4r^2}}} \lesssim \frac{|A'_\beta(r, r, d)|}{A_\beta(r, r, d)(1 - A_\beta(r, r, d))} \leq \frac{1}{r} k''_D \frac{1}{(1 - V(r))(1 - \frac{d}{2r})}$$

For a given distance d and increasing radius r, the weight $w(r, d)$ first decreases sharply but increases again once r becomes close to the maximum radius, i.e., $V(r) \approx 1$ (see Figure 2). Thus, it is high for both nodes of very low and very high radius. Clearly, the presence of a low-radius common neighbor gives strong evidence that d is small. On the other hand, the *absence* of a *very* high degree node gives strong evidence that d is very large. Note that, the presence of low radius common neighbors in the absence of very high radius common neighbors is extremely unlikely. This is because, if a pair of nodes are close enough to connect to a low radius node, they are also very likely to both be within the radius of some very high radius node.

Since $NV(r)$ is the expectation of the indegree of a node of radius r, such high-radius nodes are expected to have extremely high degrees. However, high-degree nodes in real-world settings typically connect to no more than $10 - 20\%$ of the set of nodes, which is why a practical weighting only needs to focus on situations where $1 - V(r) \approx 1$. For relatively small d (which are the interesting candidates for link prediction), the weights $w(r, d)$ are then well approximated by $w(r) = 1/r$ up to a constant. Note that this is identical to weighting a node by $1/(NV(r))^{1/D}$, i.e., essentially weighting a common neighbor i by $1/\deg(i)^{1/D}$.

Now we will discuss a popular weighting scheme, namely Adamic/Adar, which weights the common neighbors by $1/\log(\deg(i))$.

Adamic/Adar: A popular link prediction heuristic. In practice, the radius of a node is analogous to its degree, and hence it is natural to weight a node more if it has lower degree. The Adamic/Adar measure (Adamic & Adar, 2003) was introduced to measure how related two home-pages are. The authors computed this by looking at common features of the webpages, and instead of computing just the number of such features, they weighted the rarer features more heavily. In our social networks context, this is equivalent to computing similarity between two nodes by computing the number of common neighbors, where each is weighted inversely by the logarithm of its degree.

$$\texttt{Adamic/Adar} = \sum_{k \in \mathcal{N}(i) \cap \mathcal{N}(j)} \frac{1}{\log(\deg(k))}$$

(Liben-Nowell & Kleinberg, 2003) have shown that this out-performs the number of common neighbors in a variety of social and citation networks, confirming the positive effect of a skewed weighting scheme that we observed in the motivating example.

We can analyze the Adamic/Adar measure as follows. In our model, the expected degree of a node k of radius r_k is simply $NV(r_k)$, so we set the weights as $w_k = 1/\log(NV(r_k))$. Let $\mathcal{S} = \sum_k w_k Y_k$, where random variable $Y_k = 1$ if k is a type-2 common neighbor of i and j, and zero otherwise. Clearly, $E[\mathcal{S}] = \sum_k w_k A(r_k, r_k, d_{ij}) = \sum_k A(r_k, r_k, d_{ij})/\log(NV(r_k))$. Let the minimum and maximum radii be $r_{\min}$ and $r_{\max}$ respectively. The following can be easily obtained from the Chernoff bound.

Lemma 4.1 $\frac{\mathcal{S}}{N}\left(1-\sqrt{\frac{3\log(NV(r_{\max}))\ln(1/\delta)}{N\cdot A(r_{\max},r_{\max},d_{ij})}}\right) \leq \frac{E[\mathcal{S}]}{N} \leq \frac{\mathcal{S}}{N}\left(1+\sqrt{\frac{3\log(NV(r_{\min}))\ln(1/\delta)}{N\cdot A(r_{\min},r_{\min},d_{ij})}}\right)$

Clearly, the error terms decay with increasing N, and for large N, we can tightly bound $E[\mathcal{S}]$. Since $E[\mathcal{S}]$ is monotonically decreasing function of d_{ij}, this translates into bounds on d_{ij} as well.

New estimators. Based on the analysis of the motivating example discussed before, we can get a lot of intuition about a general graph. We have seen that low radius common neighbors imply that distance is small, whereas fewer high degree common neighbors in the absence of any low degree common neighbors imply that distance is large. Based on these observations, we will define the following two estimators.

Consider the estimate Q_R, which is simply the fraction of nodes with radius smaller than R that are type-2 neighbors of i and j. Let N_R be the number of nodes with radius less than R. We define Q_R as follows:

$$Q_R = \frac{\sum_{r_k \leq R} Y_k}{N_R} \qquad \rightarrow \qquad E[Q_R] = \frac{\sum_{r_k \leq R} A(r_k, r_k, d_{ij})}{N_R} \leq A(R, R, d_{ij}) \tag{9}$$

Q_R is large when many low-radius nodes are type-2 common neighbors of i and j. Application of Hoeffding bounds give:

$$P\left[Q_R \leq E[Q_R] + \sqrt{\frac{1}{2N_R}\ln\left(\frac{1}{\delta}\right)}\right] \geq 1-\delta$$

We pick R so that $E[Q_R]$ is *large enough*. While iterative algorithms can give an exact value of the upper bound on d_{ij}, we will also provide a simple intuitive bound:

$$Q_R \leq A(R, R, d_{ij}) + \sqrt{\frac{1}{2N_R}\ln\left(\frac{1}{\delta}\right)} \leq V(R)\left(1-\frac{d_{ij}^2}{4R^2}\right)^{D/2} + \sqrt{\frac{1}{2N_R}\ln\left(\frac{1}{\delta}\right)}$$

$$\Rightarrow d_{ij} \leq 2R\sqrt{1-\left(\frac{Q_R - \sqrt{\ln(1/\delta)/2N_R}}{V(R)}\right)^{2/D}}$$

If we observe a large Q_R for a small R, then this upper bound gets smaller. This shows that a large number of neighbors of small degree gives a tighter upper-bound on d_{ij}. In the same spirit, we can define another estimator $T_{R'}$, such that

$$T_{R'} = \frac{\sum_{r_k \geq R'} Y_k}{N_{R'}} \qquad \rightarrow \qquad E[T_{R'}] = \frac{\sum_{r_k \geq R'} A(r_k, r_k, d_{ij})}{N_{R'}} \geq A(R', R', d_{ij})$$

A similar analysis yields a high probability lower-bound on d_{ij}:

$$T_{R'} \geq A(R', R', d_{ij}) - \sqrt{\frac{1}{2N_{R'}}\ln\left(\frac{1}{\delta}\right)} \geq V(R')\left(1-\frac{d_{ij}}{2R'}\right)^{D} - \sqrt{\frac{1}{2N_{R'}}\ln\left(\frac{1}{\delta}\right)}$$

$$\Rightarrow d_{ij} \geq 2R'\left(1-\left(\frac{T_{R'} + \sqrt{\ln(1/\delta)/2N_{R'}}}{V(R')}\right)^{1/D}\right)$$

Smaller values of $T_{R'}$ yield tighter lower bounds. This tells us that if many high degree nodes are *not* common neighbors of i and j then, we are more confident than i and j are far away.

While Q_R and $T_{R'}$ could be used with any R and R', we could also perform a sweep over the range of possible radii, computing bounds on d_{ij} for each radius using both estimators, and then retaining the best.

5 Estimators using Longer Paths in the Deterministic Model

The bounds on the distance d_{ij} described in the previous sections apply only when i and j have common neighbors. However, there will be no common neighbors if (for the undirected case) (a) $d_{ij} > 2r$, or (b) no points fall in the intersection area $A(r, r, d_{ij})$ due to small sample size N. In such cases, looking at paths of length $\ell > 2$ between i and j can yield bounds on d_{ij}. Such bounds can be useful even when common neighbors exist; in fact, we show that the mere existence of *one* common neighbor leads to stronger bounds for long paths.

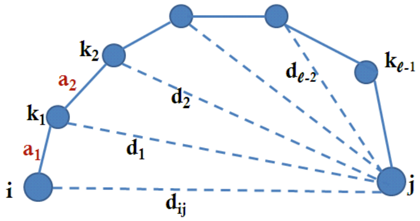

Figure 3: Triangulation for bounding d_{ij} using ℓ-hop paths.

We first discuss how d_{ij} can be upper-bounded using the observed number of simple ℓ-hop paths for any given $\ell > 2$. A stronger upper bound and a lower bound can be derived when i and j are known to have at least one common neighbor. We demonstrate this for $\ell = 3$, with a similar technique being applicable for longer paths as well. For the sake of simplicity, we restrict ourselves to the case of identical and known r.

An Upper Bound for $\ell > 2$. Let $Y(i, k_1, \dots, k_{\ell-2}, j) = 1$ if there is a simple path of length ℓ such that $i \sim k_1 \sim k_2 \sim \dots k_{\ell-2} \sim j$, with no node being repeated in the path. Let $\mathcal{S}^{(\ell)}$ be the set of ordered sets of ℓ distinct elements from $\{1, \dots, N\}$; thus, $\mathcal{S}^{(\ell)}$ represents all possible simple paths of length ℓ. Let $\eta_\ell(i, j)$ be the number of paths of length ℓ between i and j, given the distance between i, j. $\eta_\ell(i, j)$ is simply,

$$\eta_\ell(i,j) = \sum_{k_1, \dots k_{\ell-2} \in \mathcal{S}^{(\ell-2)}} Y(i, k_1, \dots, k_{\ell-2}, j | d_{ij}).$$

Note that, $\eta_\ell(i, j)$ is a function of the $N - 2$ independent random variables $X_k, k \notin \{i, j\}$, for a given $X_i = x_i, X_j = x_j$, i.e. $f_{x_i,x_j}(X_k, k \notin \{i, j\})$.

We need to infer bounds on d_{ij} given the observed number of simple paths $\eta_\ell(i, j)$. Our first step is to bound the maximum degree Δ of any graph generated by the RHH model. Next, we use Δ to bound both the maximum possible value of $\eta_\ell(i, j)$ and the change that can be induced in it by moving any one point. Finally, we compute the expected value $E(f_{x_i,x_j})$, which is simply $E[\eta_\ell(i, j)]$. Combining these will give us a bound linking d_{ij} to the number of simple ℓ-hop paths.

Under the assumption that the positions X_k are uniformly distribution in the unit hypersphere, we can bound Δ, as follows:

Lemma 5.1 $\Delta < NV\left(1 + \sqrt{\frac{\ln(N/\delta)}{2NV}}\right)$ *with probability at least* $1 - \delta$.

Proof: The degree $\deg(k)$ of any node k is a binomial random variable with expectation $E[d(k)] = NV$, where V is the volume of a hypersphere of radius r. Thus, using the Chernoff bound, $d(k) < NV\left(1 + \sqrt{\frac{\ln(N/\delta)}{2NV}}\right)$ holds with probability at least $(1 - \delta/N)$. Applying the union bound on all nodes yields the desired proposition. ■

Lemma 5.2 *For any graph with maximum degree* Δ*, we have:* $\eta_\ell(i, j) \leq \Delta^{\ell-1}$.

Proof: This can be proved using a simple inductive argument. If the graph is represented by adjacency matrix $\mathbf{M}$, then the number of length ℓ paths between i and j is given by $\mathbf{M}^\ell(i, j)$. Trivially $\mathbf{M}^2_{ij}$ can be at most Δ. This happens when both i and j have degree Δ, and their neighbors form a perfect matching. Assuming this is true for all $m < \ell$, we have: $\mathbf{M}^\ell(i, j) = \sum_p \mathbf{M}(i, p)\mathbf{M}^{\ell-1}(p, j) \leq \Delta^{\ell-2} \sum_p \mathbf{M}(i, p) \leq \Delta^{\ell-1}$ ■

Lemma 5.3 *For* $\ell < \Delta$, $|\eta_\ell(i, j|X_1, \dots, X_p, \dots, X_N) - \eta_\ell(i, j|X_1, \dots, \tilde{X}_p, \dots X_N)| \leq (\ell - 1) \cdot \Delta^{\ell-2}$

Proof: The largest change in $\eta_\ell(.)$ occurs when node p was originally unconnected to any other node, and is moved to a position where it can maximally add to the number of ℓ-hop paths between i and j (or vice versa). Consider all paths where p is m hops from i (and hence $\ell - m$ hops from j. From Lemma 5.2, the number of such paths can be at most $\Delta^{m-1} \cdot \Delta^{\ell-m-1} = \Delta^{\ell-2}$. Since $m \in \{1, \dots, \ell - 1\}$, the maximum change is $(\ell - 1) \cdot \Delta^{\ell-2}$. ■

The bounds in both Lemma 5.2 and 5.3 are tight, as can be seen by considering a clique of i, j, and $\Delta - 1$ other nodes. Next, we compute the expected number of ℓ-hop paths. Define $A(r_1, r_2, d)$ as the volume of intersection of two balls of radii r_1 and r_2, whose centers are distance d apart (as always, the dimension D is implicit). Define $Pr_\ell(i, j)$ as the probability of observing an ℓ-hop path between points i and j.

Theorem 5.4 $E[\eta_\ell(i,j)] \le \Delta^{\ell-1}\prod_{p=1}^{\ell-1} A(r, p\times r, (d_{ij}-(\ell-p-1)r)_+)$, *where* x_+ *is defined as* $\max(x,0)$.

Proof: Consider an ℓ-hop path between nodes i and j as in figure 3. The solid lines are the edges in the path, whereas the dotted lines are the distance variables we will introduce in order to compute the probability of this path. For clarity of notation, let us denote the distances d_{ik_1}, d_{ik_2} etc. by a_1, a_2, up to $a_{\ell-1}$. These are all less than r, since the corresponding edges are present. We also denote the distances d_{jk_1}, d_{jk_2}, etc. by d_1, d_2, up to $d_{\ell-1}$. From the triangle inequality, $d_{\ell-2} \le a_{\ell-1} + a_\ell \le 2r$, and by induction, $d_k \le (\ell-k)r$. Similarly, $d_1 \ge (d_{ij}-a_1)_+ \ge (d_{ij}-r)_+$, and by induction, $d_k \ge (d_{ij}-kr)_+$. We can now compute the probability of observing an ℓ-hop path:

$$
\begin{aligned}
&Pr_\ell(i,j) = P(i \sim k_1 \sim \ldots \sim k_{\ell-1} \sim j | d_{ij}) \\
=\;& P(a_1 \le r \cap \ldots \cap a_\ell \le r | d_{ij}) \\
=\;& \int\limits_{d_1,\ldots,d_{\ell-2}} P(a_1 \le r, \ldots, a_{\ell-1} \le r, d_1, \ldots, d_{\ell-2} | d_{ij}) \\
=\;& \int\limits_{d_1=(d_{ij}-r)_+}^{(\ell-1)r} \cdots \int\limits_{d_{\ell-2}=(d_{ij}-(\ell-2)r)_+}^{2r} P(a_{\ell-1} \le r, a_\ell \le r | d_{k-2}) P(a_{\ell-2} \le r, d_{\ell-2} | d_{\ell-3}) \ldots P(a_1 \le r, d_1 | d_{ij}) \\
\le\;& A(r, r, (d_{ij}-(\ell-2)r)_+) \times A(r, 2r, (d_{ij}-(\ell-3)r)_+) \times \ldots \times A(r, (\ell-1)r, d_{ij}) \\
\le\;& \prod_{p=1}^{\ell-1} A(r, p\times r, (d_{ij}-(\ell-p-1)r)_+)
\end{aligned}
\tag{10}
$$

Since there can be at most $\Delta^{\ell-1}$ possible paths (from Lemma 5.2), the theorem statement follows. ■

Corollary 5.5 *For* 3*-hop paths, we have:*

$$Pr_3(i,j) \le A(r,r,(d_{ij}-r)_+)) \cdot A(r,2r,d_{ij}) \qquad \textit{AND} \qquad E[\eta_3(i,j)] \le \Delta^2 \cdot A(r,r,(d_{ij}-r)_+)) \cdot A(r,2r,d_{ij})$$

Theorem 5.6

$$\eta_\ell(i,j) \le (NV)^{\ell-1}\left[\prod_{p=1}^{\ell-1} A(r, p\times r, (d_{ij}-(\ell-p-1)r)_+) + \frac{(\ell-1)\sqrt{\dfrac{1/\delta}{2}}}{\sqrt{N}V\left(1+\sqrt{\dfrac{\ln(N/\delta)}{2NV}}\right)}\right]\left(1+\sqrt{\frac{\ln(N/\delta)}{2NV}}\right)^{\ell-1}$$

with probability at least $(1-2\delta)$.

Proof: From McDiarmid's inequality (McDiarmid, 1989), we have:

$$
\begin{aligned}
&\eta_\ell(i,j) \le E[\eta_\ell(i,j)] + (\ell-1)\Delta^{\ell-2}\sqrt{\frac{N\ln(1/\delta)}{2}} \\
&\le \Delta^{\ell-2}\left[\Delta \textstyle\prod_{p=1}^{\ell-1} A(r, p\times r, (d_{ij}-(\ell-p-1)r)_+) + (\ell-1)\sqrt{\dfrac{N\ln(1/\delta)}{2}}\right] \\
&\le (NV)^{\ell-1}\left[\textstyle\prod_{p=1}^{\ell-1} A(r, p\times r, (d_{ij}-(\ell-p-1)r)_+) + \dfrac{(\ell-1)\sqrt{\frac{\ln(1/\delta)}{2}}}{\sqrt{N}V\left(1+\sqrt{\frac{\ln(N/\delta)}{2NV}}\right)}\right]\left(1+\sqrt{\tfrac{\ln(N/\delta)}{2NV}}\right)^{\ell-1}
\end{aligned}
$$

where Theorem 5.4 is applied in the step 2, and Lemma 5.1 in step 3. Note that as N increases, the second term in the summation decays, yielding tighter bounds. ■

Bounding d_{ij}. Theorem 5.6 yields an upper bound d_{ij} as follows. Only the first term in the summation depends on d_{ij}, and this term decreases monotonically with increasing d_{ij}. Thus, a simple binary search can give us the value of d_{ij} that achieves the equality in Theorem 5.6, and this is an upper bound on d_{ij}.

A looser but analytic bound can be obtained by upper-bounding all but one of the $A(.)$ terms by 1. For example, using $A(r, 2r, d_{ij}) \leq 1$ in Corollary 5.5 yields $E[\eta_3(i,j)] \leq \Delta^2 A(r, r, (d_{ij} - r)_+)$. Using this in McDiarmid's inequality yields a bound of the form

$$A(r, r, (d_{ij} - r)_+) \geq \frac{\eta_3(i,j)}{c(N,\delta)} - c'(N,\delta) \Rightarrow d_{ij} \leq r + 2r\sqrt{1 - \left(\frac{\eta_3(i,j)/c(N,\delta) - c'(N,\delta)}{V(r)}\right)^{2/D}}$$

In general, bounds for ℓ-hop paths are of the form $d_{ij} \leq \ell r(1 - g(\eta_\ell(i,j), \epsilon))$. Thus, for some $\ell' > \ell$, $\eta_{\ell'}(i,j)$ needs to be much larger than $\eta_\ell(i,j)$ for the bound using ℓ' to be stronger than that for ℓ. In particular, this shows that when enough common neighbors are present (i.e., 2-hop paths), looking at longer paths is unlikely to improve bounds and help link prediction, thus theoretically confirming the empirical observations of (Liben-Nowell & Kleinberg, 2003).

Better bounds when shorter paths exist. While the above applies in the general case, it suffers from two weaknesses: (1) the upper bound on $E[\eta_\ell(i,j)]$ (and hence on d_{ij}) derived in Theorem 5.4 gets progressively weaker as ℓ increases, and (2) we could not derive a useful lower bound on $E[\eta_\ell(i,j)]$ (the trivial lower bound is zero, which is achieved when $d_{ij} = \ell r$). Both of these can be remedied if we know that *at least* one path of length less than ℓ exists. We demonstrate the idea for $\ell = 3$, but it can be used for larger ℓ in a similar fashion. First, we prove two bounds on the probability $Pr_3(i,j)$ of observing a 3-hop path between two points whose distance is d_{ij}.

Lemma 5.7 *If there exists any 3-hop path between i and j, then, for any* $d' \in [(d_{ij} - 2r)_+, 2r]$*,*

$$A(r, d', d_{ij}) \cdot A(r, r, d') \leq \mathbf{Pr_3(i,j)} \leq [A(r, r, (d_{ij} - r)_+) - A(r, r, d')]\, A(r, d', d_{ij}) + A(r, r, d') \cdot A(r, 2r, d_{ij})$$

Proof: Consider all points that are within distance r of point i, and within a distance range $(x, x + \Delta x)$ of point j; here, Δx refers to an infinitesimal change in x. Since these points are within r of i, they can be the first hop in a 3-hop path from i to j. Also, since they are all equidistant from j, and the probability of a common neighbor between two points depends only on their distance, all of these points have the same probability $A(r, x, d_{ij})$ of forming a common neighbor with j. Let $p(r, x, d_{ij})$ denote the probability density function for such points. Triangle inequalities imply that $2r \geq x \geq (d_{ij} - r)_+$ for any 3-hop path to exist. Then,

$$\begin{aligned} Pr_3(i,j) &= \int_{(d_{ij}-r)_+}^{2r} p(r, x, d_{ij}) \cdot A(r, r, x)dx \geq \int_{(d_{ij}-r)_+}^{d'} p(r, x, d_{ij}) \cdot A(r, r, x)dx \\ &\geq A(r, r, d') \int_{(d_{ij}-r)_+}^{d'} p(r, x, d_{ij})dx \geq A(r, d', d_{ij}) \cdot A(r, r, d') \end{aligned}$$

This proves the lower-bound on $Pr_3(i,j)$. The proof for the upper bound is similar:

$$\begin{aligned} Pr_3(i,j) &= \int_{(d_{ij}-r)_+}^{d'} p(r, x, d_{ij}) \cdot A(r, r, x)dx + \int_{d'}^{2r} p(r, x, d_{ij}) \cdot A(r, r, x)dx \\ &\leq A(r, r, (d_{ij} - r)_+) \cdot A(r, d', d_{ij}) + A(r, r, d') \cdot [A(r, 2r, d_{ij}) - A(r, d', d_{ij})] \\ &\leq [A(r, r, (d_{ij} - r)_+) - A(r, r, d')]\, A(r, d', d_{ij}) + A(r, r, d') \cdot A(r, 2r, d_{ij}) \end{aligned}$$

■

The difficulty in using these bounds arises from the fact that $d' \in [(d_{ij} - r)_+, 2r]$, and d_{ij} is unknown. When we only know that *some* 3-hop path exists, then $d_{ij} \leq 3r$, and hence $d' = 2r$ is the only value of d' that is guaranteed to lie within the required interval. In fact, using $d' = 2r$ yields exactly the statement of Corollary 5.5. However, suppose we also knew that at least one 2-hop path exists. Then, $d_{ij} \leq 2r$, and so any d' in the range $r \leq d' \leq 2r$ is valid. In particular, we can use $d' = r$ to get the following bounds.

Theorem 5.8 *When* $d_{ij} \leq 2r$*, then*

$$A(r, r, d_{ij}) \cdot A(r, r, r) \leq \mathbf{Pr_3(i,j)} \leq [A(r, r, (d_{ij} - r)_+) - A(r, r, r)]\, A(r, r, d_{ij}) + A(r, r, r) \cdot A(r, 2r, d_{ij})$$

Lemma 5.9 *The upper bound in Theorem 5.8 is smaller (and hence, better) than the upper bound in Corollary 5.5.*

Proof: The proof follows from basic algebraic manipulations. ■

Thus, the presence of even one common neighbor between i and j offers better bounds using 3-hop paths. In addition, we also have a lower bound that was unavailable in the general case. These translate to sharper bounds on d_{ij}, which can be obtained via binary search as described previously. Similar results can be obtained for paths of length $\ell > 3$.

Observations. Our analysis of ℓ-hop paths yields the following observations. (1) When short paths are nonexistent or rare, the bounds on d_{ij} that we obtain through them can be loose. Longer paths can be used to yield better bounds in such cases. (2) As ℓ increases, more and more long paths need to be observed before the corresponding bound on d_{ij} becomes comparable or better than bounds obtained via shorter paths. (3) Even the *existence* of a short path can improve upper bounds obtained by all longer paths. In addition, lower bounds on d_{ij} can also be obtained. (4) The *number* of paths is important to the bound. Link prediction using the length of the shortest path ignores this information, and hence should perform relatively poorly, as observed by (Liben-Nowell & Kleinberg, 2003; Brand, 2005) and (Sarkar & Moore, 2007).

6 The Non-deterministic Case

All of the previous sections have assumed that, given the positions of points, the corresponding graph could be inferred exactly. In terms of the `RHH` model introduced in section 2, this corresponds to setting $\alpha \to \infty$. In this section, we investigate the effects of finite α. Our analysis shows that while bounds become looser, the results are still qualitatively similar.

The core idea underlying almost all of our previous results has been the computation of the probability of two nodes i and j having a common neighbor. For the deterministic case, this is simply the area of intersection of two hyperspheres, $A(r, r, d_{ij})$, for the case when all nodes have the same radius r. However, in the non-deterministic case, this probability is hard to compute exactly. Instead, we can give the following simple bounds on $Pr_2(i, j)$, which is the probability of observing a common neighbor between two nodes i and j that are distance d_{ij} apart and have identical radius r.

Theorem 6.1

$$
\begin{aligned}
Pr_2(i,j) &> \frac{1}{4}\left(A(r,r,d_{ij}) + 2e^{-\alpha d_{ij}} \cdot (V(r) - A(r,r,d_{ij}))\right) \\
Pr_2(i,j) &< \begin{cases} A(r,r,d_{ij}) + 2V(r) \cdot \dfrac{\left[1-\left(\frac{D}{\alpha r}\right)^D\right]}{\frac{\alpha r}{D}-1} & \textit{(for } \alpha r > D\textit{)} \\ A(r,r,d_{ij}) + 2D \cdot V(r) & \textit{(for } \alpha r = D\textit{)} \\ A(r,r,d_{ij}) + 2V(D/\alpha) \cdot \dfrac{\left[1-\left(\frac{\alpha r}{D}\right)^D\right]}{1-\frac{\alpha r}{D}} & \textit{(for } \alpha r < D\textit{)} \end{cases}
\end{aligned}
$$

Observations and Extensions. The importance of theorem 6.1 is that the probability of observing a common neighbor is still mostly dependent on the area of intersection of two hyperspheres, i.e. $A(r, r, d_{ij})$. However, there is a gap of a factor of 4 between the lower and upper bounds. This can still be used to obtain reasonable bounds on d_{ij} when enough common neighbors are observed. However, when we consider longer paths, the gap increases and we might no longer be able to get strong bounds.

The reason for this is that theorem 6.1 only uses the fact that probability of linking i and j is at least $1/2$ when d_{ij} is less than r. This statement is applicable to all α. However, we typically want to perform link prediction only when α is large, as small values of α yield graphs that are close to random and where no link prediction methods would work. For the case of large α, we can get much stronger lower bounds and close the factor-of-4 gap, as follows.

In order to compute the probability $Pr_2(i, j)$, we need to integrate the product of the link probabilities over the intersection of the two hyperspheres of radius r around nodes i and j. Let this region be denoted by $S(i, j)$.Suppose that, instead of integrating over $S(i, j)$, we integrate over a smaller subset $S'(i, j)$. While the volume of $S'(i, j)$ would be smaller, the minimum probabilities inside that subset could be much higher, leading to a better overall lower-bound. We consider $S'(i, j) = \{x_k | d_{ik} < r', d_{jk} < r'\}$ to be the intersection of two hyperspheres of radius $r' < r$, centered on i and j. Then, using eq. 4

$$\begin{aligned}
&Pr_2(i, j, x_k \in S'(i, j)) \\
&\geq \left(\tfrac{1}{1+e^{\alpha(r'-r)}}\right)^2 \cdot \text{vol}(S'(i, j)) \\
&\geq \left(\tfrac{1}{1+e^{\alpha(r'-r)}}\right)^2 V(r') \int_0^{\cos^{-1}\left(\frac{d_{ij}}{2r'}\right)} sin^D(t)dt \times \text{const.}_D
\end{aligned}$$

Ideally, we would like to pick r' to maximize this, but $\text{vol}(S'(i, j))$ depends on d_{ij} as well. Noting that and the effect of d_{ij} is restricted to the last term only, we propose the following formulation:

$$\text{Pick } r' \text{ to maximize } \left(\frac{1}{1+e^{\alpha(r'-r)}}\right)^2 \cdot V(r') \tag{11}$$

Lemma 6.2 *If $\alpha > D/r$, then $r' < r$.*

Thus, for large enough α, we can find a good r' which can improve the gap between upper and lower bounds of $Pr_2(i, j)$. The optimal r' gets closer to r as α increases, but its exact value has to be obtained numerically.

7 Summary and Discussion

The paper presents, to our knowledge, the first theoretical study of link prediction and the heuristics commonly used for that purpose. We formalize the link prediction problem as one of estimating distances between nodes in a latent space, where the observed graph structure provides evidence regarding the unobserved positions of nodes in this space. We present theoretical justifications of two common empirical observations: (1) the simple heuristic of counting common neighbors often outperforms more complicated heuristics, (2) a variant that weights common neighbors by the inverse of the logarithm of their degrees (Adamic & Adar, 2003) often performs better. We show that considering longer paths is useful only if shorter paths (especially, common neighbors) are not numerous enough for the bounds obtained from them to be tight enough. However, the bounds obtained from longer paths can be made significantly tighter if even a single short path is known to exist.

References

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, *25*.

Aldous, D., & Fill, J. A. (2001). *Reversible markov chains*.

Brand, M. (2005). A Random Walks Perspective on Maximizing Satisfaction and Profit. *SIAM '05*.

Faust, K. (1988). Comparison of methods for positional analysis: Structural and general equivalences. *Social Networks*.

Jeh, G., & Widom, J. (2002). Scaling personalized web search. *Stanford University Technical Report*.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*.

Kleinberg, J. (2000). The Small-World Phenomenon: An Algorithmic Perspective. *STOC*.

Liben-Nowell, D., & Kleinberg, J. (2003). The link prediction problem for social networks. *CIKM '03*.

Maurer, A., & Pontil, M. (2009). Empirical bernstein bounds and sample-variance penalization. *Conference on Learning Theory*.

McDiarmid, C. (1989). On Method of Bounded Differences. *Surveys in Combinatorics*.

McFarland, D. D., & Brown, D. J. (1973). Social distance as a metric: a systematic introduction to smallest space analysis. *Bonds of Pluralism: The Form and Substance of Urban Social Networks*.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*.

Raftery, A. E., Handcock, M. S., & Hoff, P. D. (2002). Latent space approaches to social network analysis. *J. Amer. Stat. Assoc.*, *15*, 460.

Raghavan, P. (2002). Social networks: From the web to the enterprise. *IEEE Internet Computing*.

Sarkar, P., & Moore, A. (2007). A tractable approach to finding closest truncated-commute-time neighbors in large graphs. *Proc. UAI*.

Schroeder, J., Xu, J. J., & Chen, H. (2003). Crimelink explorer: Using domain knowledge to facilitate automated crime association analysis. *ISI* (pp. 168–180).

8 Appendix

Proof of theorem 3.2. Define $\epsilon_o = \sqrt{\frac{2\mathtt{var}_N(Y_{OPT})\log 2/\delta}{N}} + \frac{7\log 2/\delta}{3(N-1)}$, $\epsilon_m = \sqrt{\frac{2\mathtt{var}_N(Y_{MAX})\log 2/\delta}{N}} + \frac{7\log 2/\delta}{3(N-1)}$, and $\epsilon_f = \epsilon_o + \epsilon_m$. We want to prove:

$$d_{OPT} \leq d_{MAX} \overset{w.h.p}{\leq} d_{OPT} + \left(\frac{2\epsilon_f}{V(1)}\right)^{1/D}$$

The lower bound follows from the definition of d_{OPT}. For the upper bound, define the function $A^{-1}(y) = d_{ij}$ s.t. $A(r, r, d_{ij}) = y$. Note that this function is well-defined since $A(.)$ is a monotonic function of d_{ij}. Using Leibniz's rule on equation 4 and some algebraic manipulation, we get

$$\frac{dA^{-1}(y)}{dy} = -\left[\frac{C_D r^D}{2r}\left(1 - \frac{\left(A^{-1}(y)\right)^2}{4r^2}\right)^{\frac{D-1}{2}}\right]^{-1}$$

Thus, the slope of $A^{-1}(y)$ is always negative, (and hence $A^{-1}(.)$ is monotonically decreasing), and has the highest magnitude when $A^{-1}(y)$ is largest, i.e., when $A^{-1}(y) = 2r$. Now,

$$\begin{aligned} d_{MAX} - d_{OPT} &= A^{-1}(A_{MAX}) - A^{-1}(A_{OPT}) \\ &\leq 2r - A^{-1}\left(A_{OPT} - A_{MAX}\right) \\ &\leq 2r - A^{-1}(\epsilon_f) \quad \text{(from Theorem 3.1)} \end{aligned} \tag{12}$$

Let $d = A^{-1}(\epsilon_f)$. Then,

$$\begin{aligned} \epsilon_f &= A(r, r, d) \geq V(r)\left(1 - \frac{d}{2r}\right)^D \\ \Rightarrow d &\geq 2r\left(1 - \left(\frac{\epsilon_f}{V(r)}\right)^{1/D}\right) \end{aligned} \tag{13}$$

Substituting eq. (13) in eq. (12) we get the desired result.

Proof of theorem 6.1. We want to prove:

$$Pr_2(i,j) > \frac{1}{4}\left(A(r, r, d_{ij}) + 2e^{-\alpha d_{ij}} \cdot (V(r) - A(r, r, d_{ij}))\right)$$

$$Pr_2(i,j) < \begin{cases} A(r, r, d_{ij}) + 2V(r) \cdot \dfrac{\left[1 - \left(\dfrac{D}{\alpha r}\right)^D\right]}{\dfrac{\alpha r}{D} - 1} & \text{(for } \alpha r > D\text{)} \\ A(r, r, d_{ij}) + 2D \cdot V(r) & \text{(for } \alpha r = D\text{)} \\ A(r, r, d_{ij}) + 2V(D/\alpha) \cdot \dfrac{\left[1 - \left(\dfrac{\alpha r}{D}\right)^D\right]}{1 - \dfrac{\alpha r}{D}} & \text{(for } \alpha r < D\text{)} \end{cases}$$

We use the following facts. When $x \leq 0$, $1/2 \leq 1/(1 + e^x) < 1$, and when $x \geq 0$, $e^{-x}/2 \leq 1/(1 + e^x) < e^{-x}$. Now, $Pr_2(i,j)$ is the probability that a randomly generated point k with position x_k links to both i and j. Let us divide the range of x_k into four parts:

(a) part $S(i,j) = \{x_k | d_{ik} \leq r, d_{jk} \leq r\}$,
(b) part $S(i,\neg j) = \{x_k | d_{ik} \leq r\} \setminus S_{i,j}$,
(c) part $S(\neg i,j) = \{x_k | d_{jk} \leq r\} \setminus S_{i,j}$,
(d) part $S(\neg i,\neg j)$ elsewhere.

We will first prove the lower bound and then the upper bound.

Lower bound on $Pr_2(i,j)$. Let $p(x_k)$ be the p.d.f (probability density function) of point k being at position x_k, and $Pr_2(i,j,x_k)$ be the probability that $i \sim k \sim j$ and that k is at position x_k. We have:

$$Pr_2(i,j,x_k \in S(i,j)) = \int_{x_k \in S(i,j)} \frac{1}{1 + e^{\alpha(d_{ik} - r)}} \frac{1}{1 + e^{\alpha(d_{jk} - r)}} p(x_k) dx_k \tag{14}$$

$$\geq \frac{1}{4}\int_{x_k \in S(i,j)} p(x_k) dx_k \geq \frac{1}{4}A(r, r, d_{ij}) \tag{15}$$

where $A(r,r,d_{ij}) = \int_{x_k \in S(i,j)} p(x_k)dx_k$ is the volume of intersections of two hyperspheres of radius r centered at i and j. The second-last step follows from both $d_{ik} - r$ and $d_{jk} - r$ being less than zero. Similarly,

$$
\begin{aligned}
& Pr_2(i,j,x_k \in S(i,\neg j)) \\
& = \int_{x_k \in S(i,\neg j)} \frac{1}{1+e^{\alpha(d_{ik}-r)}} \frac{1}{1+e^{\alpha(d_{jk}-r)}} p(x_k)dx_k \\
& > \frac{1}{4} \int_{\{x_k | d_{ik} \leq r, d_{jk} > r\}} e^{-\alpha(d_{jk}-r)} p(x_k)dx_k \\
& > \frac{1}{4} \int_{\{x_k | d_{ik} \leq r, d_{jk} > r\}} e^{-\alpha(d_{ij}+r-r)} p(x_k)dx_k \quad \text{(because } d_{jk} < d_{ik} + d_{ij} < d_{ij} + r\text{)} \\
& > \frac{1}{4} \cdot e^{-\alpha d_{ij}} \cdot (V(r) - A(r,r,d_{ij}))
\end{aligned}
$$

since $\int_{x_k \in S(i,\neg j)} p(x_k)dx_k = V(r) - A(r,r,d_{ij})$. The same formula holds for part $S(\neg i, j)$. So, even without considering part $S(\neg i, \neg j)$, we have:

$$Pr_2(i,j) > \frac{1}{4}\Big(A(r,r,d_{ij}) + 2e^{-\alpha d_{ij}} \cdot (V(r) - A(r,r,d_{ij}))\Big)$$

Upper bound on $Pr_2(i,j)$. We will use the following facts. The volume of a hypersphere of radius x is given by $V(x) = c_D x^D$ where c_D is some constant that depends on the dimension D. Hence, the probability density function for a point k being at distance x from a given point j is $\frac{d}{dx}V(x) = c_D \cdot D \cdot x^{D-1}$.

We have:

$$
\begin{aligned}
Pr_2(i,j,x_k \in S(i,j)) \quad & = \quad \int_{x_k \in S(i,j)} \frac{1}{1+e^{\alpha(d_{ik}-r)}} \frac{1}{1+e^{\alpha(d_{jk}-r)}} p(x_k)dx_k \\
& < \quad \int_{x_k \in S(i,j)} 1 \cdot p(x_k)dx_k < A(r,r,d_{ij})
\end{aligned}
\tag{16}
$$

$$
\begin{aligned}
& Pr_2(i,j,x_k \in S(i,\neg j) \cup S(\neg i, \neg j)) \\
& < \int\limits_{\{x_k | d_{jk} > r\}} \frac{1}{1+e^{\alpha(d_{jk}-r)}} p(x_k)dx_k < \int\limits_{\{x_k | d_{jk} > r\}} e^{-\alpha(d_{jk}-r)} p(x_k)dx_k \\
& < \int\limits_{d_{jk} > r} e^{-\alpha(d_{jk}-r)} c_D \cdot D \cdot (d_{jk})^{D-1} d(d_{jk}) \\
& < \frac{c_D D e^{\alpha r}}{\alpha^D} \int\limits_{y=\alpha r}^{\infty} e^{-\alpha y} y^{D-1} dy \quad \text{(using } y = \alpha d_{jk}\text{)} \\
& < \frac{c_D D e^{\alpha r}}{\alpha^D} \Gamma(D, \alpha r) \quad \text{(}\Gamma \text{ is the upper incomplete-gamma function)} \\
& < \frac{c_D D!}{\alpha^D} \sum_{k=0}^{D-1} \frac{(\alpha r)^k}{k!} \quad \text{(expansion of } \Gamma(.) \text{ function)} \\
& < \frac{c_D D^D}{\alpha^D} \sum_{k=0}^{D-1} \left(\frac{\alpha r}{D}\right)^k \quad \text{(using } D!/k! < D^{D-k}\text{)} \\
& < \begin{cases}
V(r) \cdot \dfrac{\left[1 - \left(\dfrac{D}{\alpha r}\right)^D\right]}{\dfrac{\alpha r}{D} - 1} & \text{(for } \alpha r > D\text{)} \\
D \cdot V(r) & \text{(for } \alpha r = D\text{)} \\
V(D/\alpha) \cdot \dfrac{\left[1 - \left(\dfrac{\alpha r}{D}\right)^D\right]}{1 - \dfrac{\alpha r}{D}} & \text{(for } \alpha r < D\text{)}
\end{cases}
\end{aligned}
$$

A similar formula holds for $Pr_2(i,j,x_k \in S(\neg i, j) \cup S(\neg i, \neg j))$. Summing over all of these gives an upper bound on $Pr_2(i,j)$.